THE MULTIPLE DIMENSIONS OF PUNISHMENT:
‘INTERMEDIATE’ SANCTIONS & INTERCHANGEABILITY WITH
IMPRISONMENT

by

Voula Marinos

A thesis in conformity with the requirements
for the degree of Doctor of Philosophy
Centre of Criminology.
University of Toronto

©2000 Voula Marinos
The author has granted a non-exclusive licence allowing the National Library of Canada to reproduce, loan, distribute or sell copies of this thesis in microform, paper or electronic formats.

The author retains ownership of the copyright in this thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without the author’s permission.

L’auteur a accordé une licence non exclusive permettant à la Bibliothèque nationale du Canada de reproduire, prêter, distribuer ou vendre des copies de cette thèse sous la forme de microfiche/film, de reproduction sur papier ou sur format électronique.

L’auteur conserve la propriété du droit d’auteur qui protège cette thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation.

0-612-53689-0
ABSTRACT

THE MULTIPLE DIMENSIONS OF PUNISHMENT:
'INTERMEDIATE' SANCTIONS AND INTERCHANGEABILITY WITH
IMPRISONMENT

By Voula Marinos

Doctor of Philosophy, Centre of Criminology,
University of Toronto, 2000

Reform strategies, and sentencing legislation for both adults and youth in Canada and elsewhere have attempted to create 'intermediate' sanctions that can be made equivalent and interchangeable with sentences of imprisonment. These sanctions may have the potential to reduce the use of imprisonment through interchangeability, so long as sanctions can be equivalent to imprisonment in severity. Implicit in these theories of punishment, however, is the assumption of the acceptability of interchangeable 'intermediate' sanctions. It is assumed that the most critical dimension to punishment is severity. This thesis challenges these simplistic conceptualizations of punishment, and presents an analysis of the perceived appropriateness of punishment.

I argue it is critical to examine the functions that punishments are seen (by judges and the public) as serving. Authors writing about interchangeability, and sentencing legislation for adults and youth, however, have given insufficient attention to the offence being punished and the offender who is the recipient of the punishment. It is suggested that these approaches to sentencing and punishment will be inadequate until a combination of factors are addressed that are part of the sentencing process.
The findings demonstrate that fines, community service orders, and conditional sentences cannot be assumed to be generally accepted as equally interchangeable with imprisonment across all purposes, offences, or offenders -- adults and youth. Different punishments were seen as appropriate in different contexts. The four studies presented in Chapters two through five demonstrate that denunciation is viewed as an important purpose by members of the public and judges for offences involving violence. In these cases, 'intermediate' sanctions were seen as lacking the denunciatory power that the prison achieves, and were less likely to be viewed as appropriate than imprisonment for violent and sexual offences.

It may be possible to create punishments that are seen as equivalent to imprisonment in severity. This thesis reveals, however, that there are multiple dimensions to punishment -- beyond severity -- that are considered by the public and judges when assessing the appropriateness of punishment. Punishments must be understood as being complex and varying along qualitative as well as quantitative dimensions.
ACKNOWLEDGMENTS

It is with great pleasure that I thank faculty, students, library and administrative staff at the Centre of Criminology. It is the consummate environment in which to learn. Open and inviting, the Centre is made up of scholars -- Professors and students -- who enjoy their work and inspire each other to challenge and be challenged. It is made up of scholars who have made my learning experience the very best that I can imagine.

I am grateful to Professor Lorna Marsden, President of York University (with the Sociology Department at the time) for guiding me to graduate school. I would also like to thank Professor Tammy Landau, Ryerson University, for guiding me to the Centre of Criminology to do a Master’s degree, where she too, experienced a fulfilling intellectual journey.

I appreciate those who helped me and supported my research interests in sentencing and punishment: Professors Donald Evans, Bill O’Grady, Diane Martin, Maeve McMahon, Jim Philips, Julian Roberts, Philip Stenning, and Dr. Carol LaPrairie, Department of Justice Canada, and Barbara Hill & Bill Sparks of the John Howard Society of Ontario. Special thanks goes to Judge David Cole of the Ontario Court of Justice, Scarborough Court. He has supplied me with case after case on conditional sentences over the last four years, and has always included me within an intellectual community committed to sentencing issues in Canada. I thank him for his ongoing support and interest in my work, and of course, for his encouragement to “just finish”!

Each person’s faith in my work has meant a great deal.

I would also like to thank the Graduate Directors during my years at the Centre, including Richard Ericson, Clifford Shearing, and Rosemary Gartner, for continually providing a fun and scholarly environment in which to work and learn. Faculty members and staff at the Centre -- past and present -- have also been helpful in the exchange of ideas and support. It is a pleasure to thank Professors Rosemary Gartner, Clifford Shearing, Carolyn Strange, Mariana Valverde, and Scot Wortley for providing me with intellectual inspiration and help. Thanks to Professor Philip Stenning who provided me with opportunities to develop a range of criminological interests through research assistantships. I appreciate my committee members --Professors Kent Roach and Clifford Shearing -- for contributing to my project through the development of ideas, their research, and for their quick reviews of drafts of my dissertation.

I would like to thank staff at the Centre -- Rita Donelan, Monica Bristol, Gloria Cernivivo and Lori Wells -- for their assistance with the practical problems of being a student, and for their friendship. Past and present Head Librarians Cathy Matthews and Tom Finlay, as well as Andrea Shier, also have assisted me in my research for many years. Thanks also to Michael Kempa and Carla Cesaroni for their speedy editorial assistance. Finally friendships with other graduate students at the Centre have provided me with years of encouragement, support, inspiration and a ton of laughs. Many thanks to Myrna Dawson, Willem De Lint, Simon Devereux, Greg Smith, Kelly Hannah-Moffat, Kimberly White-Mair, Jen Wood, and Mary-Lynn Young. It was fun to share my last year with Carla Cesaroni. Dawn Moore, and Renisa Mawani. To my two “partners in crime” --
Professors Jane Sprott and Kim Varma -- I appreciate their ongoing friendship and would like to thank them for a fabulous experience in grad school.

Some of this research on public and judicial attitudes about sentencing has been made possible through the support of Marg Stanowski, Operation Springboard, and David Daubney and Carol LaPrairie, Department of Justice Canada. Funding for this research has also been provided, in part, by the Social Sciences Research Council of Canada through a grant to A.N. Doob, doctoral fellowship to V. Marinos, as well by an Ontario Graduate Scholarship, University of Toronto Open Fellowship, and Hellenic-Canadian Federation of Ontario Educational Scholarship. I would also like to thank William Bateman, Woodsworth College, for encouraging graduate students to learn through experience as a teaching assistant.

My interest in finding ways to decrease the use of imprisonment, while also understanding the cultural limits to some strategies began in my Master's year with Professor Tony Doob. I felt encouraged to pursue research on the fine, even at times when I doubted whether there might be anything remotely interesting about a punishment involving money. His enthusiasm and dedication to research on sentencing in Canada has been an inspiration, and influenced my decision to pursue research in the field and learn from him for the coming years. As scholar, advisor, mentor and friend, I am so grateful for his ongoing support, patience and laughs. I thank him for a wonderful learning experience.

Finally I cannot thank my family enough. Thank you to my parents Nick and Christina Marinos who encouraged and supported my choices and academic pursuits, and at numerous times, were there for me beyond the call of duty. My education would have been impossible without them. I also appreciate the fact that they rarely asked "So, when will you be finished?" Thanks also to my brother and sister, Spiro and Georgia and their partners, for years of encouragement to hang on. I also have been lucky to share the joys and frustrations of writing a dissertation with my partner, friend and family, John Pappas. I thank him for his editing, flow charts, countless deliveries of home-made food to my office, playing piano over the telephone, and most importantly his love and support. His sensibility about balancing work and life have been extremely helpful and exciting.
TABLE OF CONTENTS

Chapter 1: Neglected Features of Sentencing and Punishment: Multidimensionality and the Role of the offence ................................................................. 1

Chapter 2: Public Sensibilities About Interchangeability: Fines and Community Service Orders .................................................................................. 31

Chapter 3: Public Sensibilities About Conditional Sentences of Imprisonment .......... 67


Chapter 5: Judicial Sensibilities Toward Conditional Sentences of Imprisonment and the Denunciatory Function of punishment .............................................. 104

Chapter 6: Multiple Dimensions of Punishment and the Role of the Offence: Implications for Sentencing and Punishment .................................................. 161

References: .................................................................................................................. 171
List of Tables

Chapter 2:

Table 2.1: Percent of Respondents Willing to Substitute a Fine or a Community Service Order For a Thirty Day Prison Sentence For Three Different Offences For an Adult Offender...........................................................................................................36

Table 2.2: Percent of Respondents Willing to Substitute a Fine or Community Service Order For Three Different Offences for a Young Offender.................................................................39

Table 2.3: Percent of Respondents Willing to Substitute a Fine for the Offences of Stealing Cds, Assault, Sexual Assault, and Pooled Across Three Offences, for Adult and Young Offenders...............................................................40

Table 2.4: Percent of Respondents Willing to Substitute a Community Service Order for the Offences of Stealing Cds, Assault, Sexual Assault, and Pooled Across Three Offences, for Adult and Young Offenders...............................................................40

Table 2.5: Amount of Fine Chosen as Substitute for a 30 Day Sentence of Imprisonment for Theft, Assault, and Sexual Assault Committed by Adult Offender (by Those Willing to Substitute)..................................................................................................42

Table 2.6: Number of Hours of Community Service Order Chosen as Substitute for a 30 Day Sentence for Theft, Assault, and Sexual Assault Committed by Adult Offender (By Those Willing to Substitute)..................................................................................................43

Table 2.7: Amount of Fine Chosen as Substitute for a 30 Day Prison Sentence for Theft, Assault, and Sexual Assault Committed by Young Offender (by Those Willing to Substitute)...............................................................................................43

Table 2.8: Number of Hours of Community Service Chosen as Substitute for a 30 Day Prison Sentence for Theft, Assault, and Sexual Assault Committed by Young Offender (By Those Willing to Substitute)...............................................................................................44

Table 2.9: Respondents’ Willingness to Support Interchangeability of a Community Service Order for a Sentence of Imprisonment and Their Perceptions of Completion Rates for Community Service Orders for Adult Offenders (Pooled Offences).............................................................................................................46

Table 2.10: Respondents’ Willingness to Support Interchangeability of a Community Service Order for a Sentence of Imprisonment and Their Perceptions of Completion Rates for Community Service Orders for Young Offenders (Pooled Offences).............................................................................................................46
Table 2.11: Willingness to Support Interchangeability of a Community Service Order for a Sentence of Imprisonment and Perceptions of Government Investing in Building More Prisons or Alternatives to Imprisonment for Adult Offenders (Across Pooled Offences) ........................................................................48

Table 2.12: Willingness to Support Interchangeability of a Community Service Order for a Sentence of Imprisonment and Perceptions of Government Investing in Building More Prisons or Alternatives to Imprisonment for Young Offenders (Pooled Offences) ........................................................................49

Table 2.13: Willingness to Support Interchangeability of a Community Service Order for a Sentence of Imprisonment and Perceptions of Government Investing in Building More Prisons or Crime Prevention for Adult Offenders (Pooled Offences) .................49

Table 2.14: Willingness to Support Interchangeability of a Community Service Order for a Sentence of Imprisonment and Perceptions on Government Investing in Building More Prisons or Crime Prevention for Young Offenders (Pooled Offences) ........................................................................50

Table 2.15: Respondents’ Views on the Importance of Denunciation as a Sentencing Purpose and Their Willingness to Support Interchangeability of a Fine for Adult Offenders ........................................................................53

Table 2.16: Respondents’ Views on the Importance of Denunciation as a Sentencing Purpose and Their Willingness to Support Interchangeability of a Community Service Order for Adult Offenders ........................................................................53

Chapter 3

Table 3.1: Preference for Prison or Conditional Sentence as a Function of Offence ........................................................................71

Table 3.2: Appropriateness of a Sentence of Prison, Conditional Sentence, or Probation for a Break and Enter into a Commercial Establishment by a Man with a Criminal Record ........................................................................73

Table 3.3: Preference for a Normal Prison Sentence or a Conditional Sentence as a Function of Respondents’ Views of the Severity of Adult Sentences ........................................................................75

Table 3.4: Preference for a Normal Prison Sentence or a Conditional Sentence as a Function of Respondents’ Preference for More Prisons or More Use of Alternatives to Custody ........................................................................76

viii
Table 3.5: Preference for a Normal Prison Sentence or a Conditional Sentence as a Function of Respondents' Preference for More Prisons or Investment in Crime Prevention

Chapter 4

Table 4.1: The Use of ‘Intermediate’ Sanctions, Alone and In Combination with Other Sanctions, for Selected Offences

Table 4.2: The Use of ‘Intermediate’ Sanctions Alone and In Combination with Prison and Probation

Table 4.3: The Proportion of those Cases that Received a Short, Medium or Long Custodial Sentence That Also Received an ‘Intermediate’ Sanction

Chapter 5

Table 5.1: Respondents’ Views About a Conditional Sentence Accomplishing the Purposes of Sentencing as Effectively as Imprisonment

Table 5.2: Judges’ Views About the Effectiveness of a Conditional Sentence in Achieving Denunciation and Their Views About Public Understanding of The Sanction

Table 5.3: Judges’ Views About the Effectiveness of a Conditional Sentence in Achieving Denunciation and Their Views About Public Support for Use of Sanction

Table 5.4: Judges’ Views About the Ability of a Conditional Sentence to Accomplish Denunciation and Their Consideration of The Impact of a Conditional Sentence on Public Opinion

Table 5.5: Judges’ Views About the Effectiveness of a Conditional Sentence in Achieving Rehabilitation and Their Views About Public Understanding of The Sanction

Table 5.6: Judges’ Views About the Effectiveness of a Conditional Sentence in Achieving Rehabilitation and Their Views About Public Support for Use of Sanction

Table 5.7: Judges’ Views About the Ability of a Conditional Sentence to Accomplish Rehabilitation and Their Consideration of The Impact of a Conditional Sentence on Public Opinion
Table 5.8: Correlation Matrix of Judges’ Perceptions of the Ability of the Conditional Sentence of Imprisonment to Achieve the Purposes of Sentencing ...............124

Table 5.9: Judges’ Views About the Ability of a Conditional Sentence to be as Effective (With Conditions) as Imprisonment and Their Views on the Public’s Understanding of the Sanction .................................................................126

Table 5.10: Judges’ Views About the Ability of a Conditional Sentence to be as Effective (With Conditions) as Imprisonment and Their Views on Public Support of the Use of the Sanction .................................................................126

Table 5.11: Judges’ Views About the Ability of a Conditional Sentence to be as Effective (With Conditions) as Imprisonment and Their Consideration of the Impact of the Sanction on the Public .................................................................127

Table 5.12: Judges’ Use of Conditional Sentences and Their Views That Each Purpose of Sentencing Can be Achieved (With Conditions) as Effectively as Prison “Always and Usually” .................................................................128

Table 5.13: Judges’ Views About the Ability of a Conditional Sentence to be as Effective (With Conditions) as Prison and Their Perceptions That the Public Can be Made to Understand the Difference Between the Conditional Sentence and Probation .................................................................133

Table 5.14: Judges’ Views About the Ability of the Conditional Sentence to be as Effective (with Conditions) as Prison and their Perceptions That the Public Believe That a Conditional Sentence Has a Different Impact on Offenders Than a Probation Order With Same Conditions .................................................................133

Table 5.15: Judges’ Views About the Ability of a Conditional Sentence to be as Effective (With Conditions) as Prison and Their Views on Denying Bail for Breaching .................................................................136

Table 5.16: Judges’ Views About the Ability of a Conditional Sentence to be as Effective (With Conditions) as Prison and Their Views on Automatic Prison for Breaches .................................................................137

Table 5.17: Judges’ Views About Denunciation and Their Attitudes Towards the Unsuitability of Violent Offences for Conditional Sentences .................................................................141

Table 5.18: Judges’ Views About Denunciation and Their Attitudes Towards the Unsuitability of Property Offences for Conditional Sentences .................................................................142

x
List of Appendices

Chapter 2: Appendix A – Respondents’ Views on the Importance of Each Sentencing Purpose (Deterrence, Incapacitation, Rehabilitation, and Compensation) and Their Willingness to Support Interchangeability of a Fine or CSO for Adult Offenders....................................................................................59

Chapter 2: Appendix B – Respondents’ Views on the Importance of Each Sentencing Purpose (Deterrence, Incapacitation, Rehabilitation, and Compensation) and Their Willingness to Support Interchangeability of a Fine or CSO for Young Offenders........................................................................................................63

Chapter 5: Appendix A – Interview with Court of Appeal Judges on Conditional Sentences........................................................................................................................................157
PREFACE


---

1 The University of Toronto School of Graduate Studies 1999-2000 Regulations now provide for the possibility that a thesis “may contain a selection of several papers”. See 1999-2000 Calendar, at p. 50, Degree Regulations, Doctor of Philosophy.
Chapter 1

Neglected Features of Sentencing and Punishment: Multidimensionality and the Role of the Offence

Attempts to reduce the use of imprisonment for both adult and young offenders have occupied reformers in a number of countries. The approaches to this perceived problem have varied. For example in (previously West) Germany, legislation was introduced in the 1970s to substitute fines for sentences of imprisonment of less than six months, as well as create a day-fine system that allowed offenders to pay a fine in exchange for dismissed charges by prosecutors (Weigend 1997). In England and Wales, a successful extension of community service orders and the opening of day reporting centers apparently contributed to a decrease in the use of imprisonment (Tonry 1995: 175). Finally, some approaches appear to have successful results in some countries and be less effective in others. For example, “truth in sentencing” legislation was introduced in Victoria without increasing prison populations, while an increase in prisoners occurred in New South Wales, Australia (Freiberg 1997; Tonry 1995: 175).

In Canada, similar concerns have been expressed for over twenty-five years. In fact in 1969, the federally appointed Canadian Committee on Corrections recommended an increased use of alternatives to imprisonment in order to decrease Canada’s prison population (1969: 309). In a comprehensive review of Canada’s sentencing system, The Canadian Sentencing Commission concluded that “[w]hat is now needed is not further theoretical development of this concept but a policy which transforms the principle of restraint into reality” (1987: 46). One of the main thrusts (principle 9) of the “Daubney” Report in 1988 was that “carceral sentences should be used with restraint; there must be a
greater use of community alternatives to incarceration…” (Standing Committee on Justice and Solicitor General, 1988: 5, hereafter “Daubney”).

One common approach to reducing the use of imprisonment is to focus on non-prison sanctions. These are traditionally, or commonly, referred to as ‘community-based sanctions’ (Canadian Sentencing Commission 1987), and ‘alternatives to imprisonment’ (McMahon 1992). Instead of these terms, authors Morris and Tonry (1990) refer to these as “intermediate punishments” -- punishments that lie between probation (seen as little or no punishment) and imprisonment (with the exception of capital punishment, seen as the most severe punishment available) (see also Daubney 1988). The two ends of the continuum -- probation and prison -- are seen by the authors as overused punishments (Morris and Tonry 1990: 3-8). The term ‘intermediate’ sanctions attempts to move away from a focus on either the prison or probation, and move towards highlighting the range of sanctions in between.

Morris and Tonry (1990) have encouraged the substitution or interchangeability of ‘intermediate’ sanctions that are equivalent to imprisonment. Equivalency, they suggest, could be devised for most short or intermediate length prison sentences. Indeed, they suggest that ‘intermediate’ punishments -- fines or community service orders for example -- could act as “roughly equivalent” substitutes for sentences of up to two years in prison (1990: 79). Offenders in the U.S. are sentenced to prison of two years or less for a range of property and relatively minor violent offences (Morris and Tonry 1990: 79).

In Canada, legislation currently exists which appears to have as its intent the reduction in the use of imprisonment. Two specific provisions of the Criminal Code state that imprisonment should not be used if less restrictive sanctions are appropriate (section
718.2(d)), and that imprisonment should not be used unless all other available sanctions have been considered (section 718.2(e)). The overall “fundamental” principle of sentencing for adults, under section 718.1, is proportionality: “[A] sentence must be proportionate to the gravity of the offence and the degree of responsibility of the offender”.

The *Young Offenders Act* includes similar restrictions on the use of imprisonment. In deciding whether custody is to be imposed, the court is directed to consider non-custodial dispositions when appropriate for young people who commit an offence that “does not involve serious personal injury” (section 24 (1.1)(b)). The courts are also to impose custody only “when all available alternatives to custody that are reasonable in the circumstances have been considered” (section 24 (1.1)(c)). In addition, proportionality in sentencing young persons is also an important principle. One difference between the principle of proportionality for young people, compared to adults, is that in the context of imposing a custodial sentence the youth court should consider the “seriousness of the offence and the circumstances in which it was committed and having regard to the needs and circumstances of the young person” (section 24 (1.1), *Young Offenders Act*).

Clearly sentencing legislation in Canada has been constructed, in recent years, with the underlying purpose of limiting the use of imprisonment and encouraging the use of non-custodial sanctions available by the courts for both adult and young offenders (Campbell 1999: 138). Although the principles in the two pieces of legislation (*Criminal

---


2 In the Canadian context, both policy documents and legal changes reinforce intermediacy in punishment as a strategy to reduce the use of imprisonment. In 1991, the Sentencing Reform Team, Department of Justice Canada, provided a background paper and proposal for changes to provisions relating to probation, conditional discharges, suspended sentences, and others. Within this document, the changes proposed were described in the context of “intermediate” sentencing, and the need to decrease the use of imprisonment (1991: 4-5). *Bill C-90 (1992) recommended the increased use of “intermediate” sanctions, but the proposed
Code and the YOA) differ to some extent, the overall intent is the same. It is suggested, in effect, that judges should substitute an ‘intermediate’ sanction for sentences of imprisonment that might otherwise be imposed.

The difficulty is that these implicit theories of punishment assume notions of the general acceptability of interchangeability. Morris and Tonry (1990) assume that a range of different punishments will be seen as appropriate by members of the public and judges in accomplishing a range of different purposes for adults and youth. These two key pieces of sentencing legislation for adults and youth also do not consider the offence that is the basis of the punishment, characteristics of the accused, or the purpose that the sentence is supposed to accomplish in a comprehensive way.

This thesis suggests that these approaches to understanding punishment in the context of sentencing in the criminal justice system will be inadequate until other parts of the punishment process are addressed and incorporated. It is possible, for instance, that a sentence of imprisonment might be seen as a necessary response in a particular case even though a sentence of “equivalent” severity -- perhaps a fine or a community service order -- could be devised. In this thesis, both members of the public and judges are asked about their views and attitudes towards the interchangeability of fines, community service orders, and conditional sentences for sentences of imprisonment. This thesis will present data that suggest that the perceived appropriateness of a punishment is determined by more than the relationship of the seriousness of the offence and the severity of the sanction. Rather I suggest that in order to understand whether a particular sentence is
perceived as appropriate, it is necessary to examine the functions that sentences are seen as serving -- by judges and members of the public -- as well as the characteristics and qualities of the offence and the offender.

The ‘Functions’ of ‘Punishment’

Writings on punishment can be summarized under at least four separate headings. The first might be described as attempting to understand the broad functions of state imposed punishment. These functions may or may not be explicit and intended.3 In the criminal justice context, the views of Cesare Beccaria (1767/1971) and Immanuel Kant (1785/1972) are often contrasted as representing the two opposing yet dominant justifications for punishment today.

Beccaria is typically associated with a classical view, where punishments are imposed for a very specific purpose: to deter crime. The focus is on the future -- preventing crimes that have not yet occurred. At the same time, however, Beccaria endorsed a form of proportionality in punishment. The severity of the punishment should not exceed that which is necessary in order to deter. Its purpose is to improve society. But for Beccaria, all punishments were not interchangeable. Given its purpose, the ideal punishment was one that would remind potential offenders of the offence and deter them from committing such offences (Beccaria 1767/1971).

3My intention in this discussion is not to suggest that punishment should only be examined in terms of intended purposes. Unintended consequences also should not be interpreted as serving intended purposes. On a broad level, Hudson (1996) explains that "sociological functionalism" assumes that all social institutions must have a function in order to survive (1996: 133). I would like to borrow Duff’s and Garland’s (1994) conception of ‘latent’ functions of punishment. They mean "...identifying the ways in which penal practice may come to serve ends other than those that are officially declared as its objectives" (Duff and Garland 1994: 32).
The other dominant approach in today's penal culture is derived from Kant. His perspective was that punishment must not be imposed on people for the good of others. Rather punishment should be imposed on an offender simply because that person has offended (*jus telionis*) -- as an end in itself. The choice of the exact punishment and the quantity of it, in general, was to be determined by the wrong that was actually committed. The principle of equality -- that punishment should reflect the crime in quality and quantity -- was the starting point. Kant realized, however, that sometimes the impact of the punishment would not be appropriate if this strict equivalence were followed (1972: 103-4). He states that social status of the offender, for example, has an impact on the pain of punishment:

It may appear, however, that difference of social status would not admit the application of the principle of retaliation, which is that of 'like with like'. But although the application may not in all cases be possible according to the letter, yet as regards the effect it may always be attained in practice, by due regard being given to the disposition and sentiment of the parties in the higher social sphere. Thus a pecuniary penalty on account of a verbal injury, may have no direct proportion to the injustice of slander; for one who is wealthy may be able to indulge himself in this offence for his own gratification. Yet the attack committed on the honour of the party aggrieved may have its equivalent in the pain inflicted upon the pride of the aggressor, especially if he is condemned by the judgment of the court, not only to retract and apologize, but to submit to some meaner ordeal, as kissing the hand of the injured person. In like manner, if a man of the highest rank as violently assaulted an innocent citizen of the lower orders, he may be condemned not only to apologize but to undergo a solitary and painful imprisonment, whereby, in addition to the discomfort endured, the vanity of the offender would be painfully affected, and the very shame of his position would constitute an adequate retaliation after the principle of like with like (1972: 105).

Kant would appear to have a notion of interchangeability where "substitute" punishments would be imposed largely because the pain inflicted on society by the offender would not have the same impact if inflicted on the offender by society.
More broadly, Durkheim’s *Moral Education* (1961) demonstrated that institutional punishment operates beyond the criminal justice system. For Durkheim, discipline is socially useful and necessary in the classroom, for instance, in order to reinforce proper behaviour, relationships and moral communication (1961: 37). He states that

Thus the essential function of punishment is not to make the guilty expiate his crime through suffering or to intimidate possible imitators through threats, but to buttress those consciences which violations of a rule can and must necessarily disturb in their faith – even though they themselves aren’t aware of it; … and, to speak more particularly of the school situation, that it is always felt by the teacher from whom the children receive it. Thus, discipline plays an important part in the functioning of the morality of the school.

Pain, however, is only an incidental repercussion of punishment; it is not its essential element. It is an external index of the feeling that must assert itself in the face of a violation. It is the feeling expressed, and the sign by which it is expressed, that neutralizes the morally disruptive effect of the violation. Thus, severity of treatment is justified only to the extent that it is necessary to make disapproval of the act utterly unequivocal.

In contrast to Kant, then, Durkheim would suggest that interchangeable sanctions are permissible only in so far as the severity of punishments are not only equivalent in severity, but more importantly, that the nature of punishments are equivalent in the messages that are communicated.

More recently, the focus of discussions about the functions of punishment has shifted in two ways. First, other functions of punishment have been discussed within broad social and economic terms. For example, in *Punishment and Social Structure*, Rusche and Kirscheimer (1939) argued that in addition to penal objectives, punishment should be understood within the context of social relations at the time. Drawing on Marxist theory, Rusche and Kirscheimer demonstrated that punishment and the labour market are
interrelated. Different punishments were structured primarily, although not exclusively, by economic and social forces. Punishment is understood as functioning to support dominant class economic interests and relations, and "...was not the result of humanitarian considerations" (1939: 24).

Second, state punishment in a criminal justice context has been linked, more generally, to state control. For example, Stanley Cohen (1985) suggested that alternatives to imprisonment were ineffective in decarceration, extended state social control to newer groups of deviants, and intensified controls directed at former deviants (at 40; see Chapter 2; see also Cohen and Scull 1983). Whether this is generally the case with alternatives to imprisonment is questioned by McMahon (1990). She found that in the context of Ontario, prison admissions declined within an approximate thirty-year period (1950s and 1980s). The development and expanded use of probation, and changing perspectives and responses to those charged with public intoxication led to reduction in the use of imprisonment. She concluded that decarceration is possible, but the issue must be assessed by examining 'alternatives' within the community and their impact on incarceration (McMahon 1992: 207-8).

Foucault (1977) saw criminal justice punishments as part of a much larger form of control, suggesting that 'punishment', 'correction', and 'discipline' are instrumental and operate everywhere, from the state, to the penitentiary, school, charitable societies, social agencies and so on (1977: 297-99). He analyzed, historically, the techniques of power within different institutions to understand how power dominates and subordinates. For Foucault, punishment is power. The prison is one institution that is linked to other

4 Their analysis is centered on penal developments in Europe from the Middle Ages to the industrial revolution to fascism.
institutions - a carceral network. "The 'carceral' with its many diffuse or compact forms, its institutions of supervision or constraint, of discreet surveillance and insistent coercion", Foucault argued, "assured the communication of punishments according to quality and quantity; it connected in series or disposed according to subtle divisions the minor and the serious penalties, the mild and the strict forms of treatment, bad marks and light sentences" (1977: 299). In this sense, power and coercion can be differentiated in gradations, to some extent, according to the severity and type of punishment, but along a continuum of punishments within the criminal justice system and extending beyond to other institutions.

Although I have only touched the surface of this body of literature, it would appear to be largely silent on the issue of how specific punishments are linked to purposes and to the acts that are being punished. Rather why people are, or should be punished, and how punishment is linked with other forms of state control has been the central issue within this literature.

The Purposes of Sentencing

The second set of literature -- the purposes of sentencing -- naturally flows from the first, the purposes of punishment. Within a criminal justice context, the "purposes" of sentencing are often listed as including those purposes that are consistent with earlier theorists along with a few others that have recently gained popularity. Thus, for example, Sections 718(a) through (f) of the Criminal Code provide a clear list of the purposes of sentencing, including denunciation, deterrence (general and individual), incapacitation, assisting in rehabilitation, victim reparation, and promoting accountability. Sections 730
(Absolute and Conditional Discharges) to section 752 (Declaring Dangerous and Long-Term Offenders) provide a wide range of sanctions including fines, probation, conditional sentences, restitution and so on. According to the structure of the legislation, the judge chooses the appropriate purpose or purposes to accomplish, and then chooses one or more of a range of different sanctions with which to accomplish the goal(s).

It is assumed, then, that any sanction can accomplish almost any sentencing goal. There is no explicit connection between the two within the legislation in terms of the ability of any one sanction, compared to another, to accomplish one or more purposes. Furthermore, there is no mention of the offence in this context. It is possible that the perceived ability of a sanction to accomplish a particular purpose will vary according to the nature of the offence. More broadly, the legislation is structured so that the various sanctions are not directly associated with the list of purposes and principles of sentencing that are contained in the legislation.

Other sentencing regimes that have not wanted to make difficult decisions have referred to a similar list of purposes. For example, the U.S. Sentencing Commission, created by the Sentencing Reform Act, was responsible for the development of Federal Sentencing Guidelines (1994; see Doob 1995). Wanting severity and inflexibility in sentencing (Doob 1995: 212), the Commission developed a two-dimensional grid based upon the nature of the offence (offence level) and criminal history (criminal history points) (see Miller 1999; Doob 1995). Yet many have critiqued the guidelines because of its lack of coherence in sentencing purposes (Doob 1995; Tonry 1995). The overall purpose of the sentencing process, according to the Commission, is to reduce crime, and the principles with which to guide sentencing include just punishment, deterrence,
incapacitation and rehabilitation (28 U.S.C. 991 (b) (2)). However as Doob (1995) argues, there is no mention of how these goals can be achieved most effectively with particular sanctions and according to the nature of the act, such as whether long prison sentences deter burglars more than intermediate punishments (1995: 213).

The literature on sentencing purposes stems from two opposing theories. First of all, there are arguments on whether the approach should be largely based on the offence (an approach consistent with Kant) or on preventing crime through deterrence, incapacitation, or rehabilitation. Consistent with the Classical School (Beccaria 1764; Bentham 1789), general deterrence remains as a popular purpose of sentencing. This literature is interesting, particularly from the standpoint of this thesis, because the focus is largely on the severity of the punishment rather than on the nature of the punishment itself. On occasion certain offences are discussed as being more susceptible to control through deterrence than others (Varma and Doob 1998), or that certain populations are more readily deterred than others (R. v. J.J.M. (S.C.C.)). What is not discussed within this literature generally, however, is whether there are particular punishments that might be more effective with particular offences in accomplishing each specific goal of punishment.

Similarly, the literature on incapacitation focuses largely on whether it is an efficient and cost-effective strategy to reduce crime. Advocates of incapacitation such as Greenwood and Abrahamse (1982) and Wilson (1998) argue for the crime-control benefits of incarcerating high-risk, serious offenders. However others such as Brown (1998), for example, found that selective incapacitation in New Zealand has been ineffective in reducing crime after release, unduly contains some offenders who might
not otherwise reoffend, and wastes resources (1998: 714; see also Auerhahn 1999). These studies are consistent with other literature on three strikes legislation in states such as California (e.g., Vitiello, 1997). Von Hirsch argues that selective incapacitation is "both on empirical and ethical grounds a device of limited potential, at best" (von Hirsch 1998: 126).

Incapacitation is to be achieved, then, through severe and lengthy sentences of imprisonment. What is notable about the literature on incapacitation is that the focus is almost exclusively on the prison as a means to incapacitate offenders. Seldom is the possibility that other sanctions could make it difficult for a person to re-offend and restrict his or her liberty through house arrest, attendance orders, community service orders where a person is expected to be supervised during "vulnerable" periods of the day. In fact Zimring and Hawkins (1995) suggest that while electronic monitoring and house arrest restrain offenders to some extent, they function as different forms of monitoring systems rather than systems of control such as imprisonment (1995: 157-161).

While rehabilitation in prison lost its appeal in the 1970s (Canadian Sentencing Commission 1987: 36; Cullen and Gilbert 1982), more recently many researchers focus broadly on the effectiveness of rehabilitation, within prison or through non-custodial programs. Researchers in Canada (Andrews, Zinger, Hoge, Bonta, Gendreau and Cullen 1990) undertook a meta-analysis of research on rehabilitation programs. They concluded that "[a]ppropriate correctional service appears to work better than criminal sanctions not involving rehabilitative service and better than services less consistent with our a priori principles of effective rehabilitation" (1990: 384). They suggest that there is solid
research on the effectiveness of rehabilitation (see also Whitehead and Lab 1989; Izzo and Ross 1990; Gendreau and Ross 1987). Palmer (1994) in the U.S., for example, has written extensively about research on more specific programs and interventions for youth such as group counseling/therapy, individual counseling/therapy, life skills, educational training and others, and has highlighted the outcomes of such programs in general (1994: 22-63). To date, however, research on programs is limited because there appears to be a lack of discussion of the specific character of these programs, including issues about intrusiveness, responsibility, and stigma to the offender.

The relatively recent literature on rehabilitation in the context of sentencing focuses largely on whether court-imposed rehabilitative programs can be effective. Hence, although the research has, to some extent, focused on the matching of offenders to programs, little attention has been given to the principles of punishment. In particular, the research on rehabilitative purposes of sentencing has generally ignored issues of proportionality. Achieving "just penalties" and "just deserts" (von Hirsch 1976) through proportionality, as mentioned earlier, is a fundamental principle in Canadian sentencing law (section 718.1 Code), and elsewhere. Literature by Andrew von Hirsch (1976) and others (Ashworth 1995; Morris 1998; Duff 1998) focus on the importance and process by which punishment will be seen as fair and appropriate.

For example, von Hirsch (1998) argues that what is relevant is not how an offender's punishment might affect his or her future behaviour or that of others, but the apparent fairness of the sentence to the offender relative to others with the same blameworthiness (1998: 168-73). An ordinal scale of sanctions is effective in achieving proportionality according to the severity of the offence and the harm caused. However
the appropriate sanction -- whether probation, a fine, or prison -- to achieve proportionality -- rarely is addressed.

More recently the role of the victim has become a salient feature of the criminal justice process through the development of restorative principles of sentencing (Daubney 1988; Roach 1999). Two sentencing purposes in Canada address the needs of victims -- "provide reparations for harm done to the victims or to the community" and "promote a sense of ... acknowledgment of the harm done to victims and to the community" (sections 718(e) & (f)). Second, recent amendments to the sentencing process in Canada include a verbal "victim impact" statement within court proceedings (section 722, Code). The third and related shift in sentencing, related to the inclusion of victims, is family group conferencing for young offenders (section 18, proposed YCJA). However the Criminal Code does not address how particular sanctions may be more capable of achieving these purposes than others. It is possible, for instance, that probation might serve as a more effective way, compared to a victim-fine surcharge, to promote a sense of harm done to the victim and community.

The literature in this area provides a rationale for shifting paradigms from retributive to restorative sentencing aims and results. Braithwaite (1989), Zehr (1990), and more recently Cayley (1998) argue for the merits of restorative justice approaches to sentencing and punishment (see also R. v. Gladue [1999]). Family group conferencing (LaPrairie 1995; Ontario Ministry of Attorney General 1999), sentencing circles (Roberts and LaPrairie 1996), and victim-offender mediation programs (Rudin 1999) have been developed consistently with the principles of restorative justice.
However, some critiques have suggested that the traditional justice system is not the appropriate setting to accomplish these goals for victims (Goddu 1993), and that victims' rights reinforce crime control (Roach 1999: 279). Others have addressed the difficulties in achieving restorative justice in the criminal justice process for female and aboriginal victims in particular (see Roberts and LaPrairie 1996; Roach 1999), as well as possible effects of sentencing disparity among offenders. Apart from discussions about the victim-fine surcharge and restitution (i.e., Ashworth 1986; Roach, 1999: Chapter 9), there is less literature about how to accomplish victim-related purposes through traditional justice punishments. Sentences of imprisonment, fines, and probation, for instance, remain focused largely upon the offender.

As mentioned earlier, the Criminal Code presents a list of purposes of sentencing from which the sentencing judge is to choose one or more appropriate purposes. In addition, the Code contains a statement that sentences must be proportionate to the harm done (Section 718.1). Sections 730 to 752 present a list of punishments, most of which can be imposed on almost any offender. The major exceptions are absolute and conditional discharges (section 730), in which the “offender” cannot have been convicted of a serious offence (punishable by fourteen years or more). What is implied by this structure -- separate lists of offences, purposes of sentencing, and sanctions or punishments -- is that sentences can be crafted for any offence to accomplish more or less any purpose with almost any sanction. There is, therefore, an implication that sanctions are largely interchangeable and can be imposed to accomplish a wide range of purposes across various offences.
**Shifts in the Justification of Punishment**

Another set of literature outlines how shifts in punishment have taken place. Over time, it is possible to point to a number of underlying reasons or justifications for the imposition of punishment. For example, in the past fifty years one can point to changes in the importance of rehabilitation, as well as the development of a fairly recent justification of punishment – identifying and managing high-risk offenders. These justifications of punishment focus largely on how both the prison and sanctions within the community accomplish these goals of sentencing. These literatures are less likely to address, however, a discussion about the possibility that some punishments, compared to others, might be seen as necessary and justified for reasons related to the nature of the offence being addressed, and characteristics of the offender (youth or adult).

From the turn of the nineteenth century until the 1970s, rehabilitating offenders was a major aim and justification of punishment within the United States and Canada. Underlying the justification of rehabilitation were positivist conceptions about crime: that there are biological and psychological causes of crime that can be addressed, and each person engages in crime for different reasons (Cullen and Gilbert 1982: 10-11). Before the late 1800s, punishment was imposed predominately for the sake of retribution and deterrence and sentencing was largely a determinate and fixed system (Canadian Sentencing Commission 1987: 36-37; Cressey 1982: xvii).

With the shift to rehabilitation, offenders remained in prison until they were "cured". The difficulty in predicting recovery or rehabilitation of offenders required the courts to impose indeterminate sentences (Cullen and Gilbert 1982: 11). It was assumed that rehabilitation could "work" effectively within prisons just as it did within hospitals.
The apparent interchangeability of curing offenders within the hospital and prison settings, however, raised some difficulties. In some instances, "milieu therapy" was required, where attempts were made to make "guards" and "treatment staff" indistinguishable within prisons (Martinson 1974). This appears to be a way of 'taking out' the prison from correctional rehabilitation, and create a more supportive and therapeutic environment. Over time, however, rehabilitation became to be viewed as a "failure".

In 1974, Martinson set out to review the literature and research studies on rehabilitation within prisons to answer the question "What works?". He concluded in a published paper that rehabilitation was not effective in reducing recidivism. While Martinson was not highly optimistic about the future of rehabilitation, he provided numerous possible reasons for this conclusion as well as the need for further studies. One quote within the paper, however, was interpreted by many to mean that "nothing works": "With few and isolated exceptions, the rehabilitative efforts that have been reported so far have had no appreciable effect on recidivism" (1974: 25). In conclusion, Martinson suggested that while rehabilitation is based upon the notion that crime is a "disease" to be cured, a perspective that focuses on deterring and preventing crime through punishment might be more effective (Martinson 1974: 50). It is likely that rather than the single purpose of rehabilitation, the prison was also seen as being required for punishment and restraint for persons convicted of serious violent and sexual offences.

Largely as a response to the dissatisfaction with rehabilitation (see Allen 1998; Cullen and Gilbert 1982; von Hirsch 1976), desert in sentencing developed prominence in discussions of sentencing in works such as von Hirsch's Doing Justice (1976). 'Just
deserts’ has been characterized by a focus on ‘just’ penalties. The second major shift in justifications of punishment addressed within the literature has been from a logic of ‘criminal justice’ in modern society, to a logic of ‘risk’ in postmodern society (Shearing 1997: 5-7). The focus within this literature appears to be how the justification of risk has determined the ways in which a range of sanctions function and are seen as appropriate.

At a broad level, Ulrich Beck (1992) provided one of the earliest accounts of the rationale of ‘risk’ within contemporary societies in the early 1990s. He argued that postmodern society is a risk-oriented society. He stated that there has been a shift from problems relating to the distribution of wealth in modern society, to a period of advanced modernization in postmodern society that is characterized by the distribution of risk. “We are living in the period of transition,” Beck states, “in which the problems of distribution of wealth and of risks overlap” (1992: 200). He presents evidence of this shift within the labour market, traditional institutions such as the family and marriage, science, and dangers within the environment such as dying forests. Lowering probabilities of harm to these institutions and systems of ecology — implying a negative standpoint — becomes a way to decrease uncertainty and risk (1992: 208).

A number of authors have identified how the logic of ‘risk’ plays out or how populations are governed through risk within criminal justice (O’Malley 1996; Shearing 1997; Ericson and Haggerty 1997). Feeley and Simon (1992) in particular, analyze how the shifts in the mentalities, institutions and practices of punishment from the ‘old’ to the ‘new penology’ have been explained largely by a single justification of punishment. They argue that whereas the “old penology” was concerned with rehabilitating and
reforming offenders, the new penology is less concerned “with responsibility, fault,...intervention and treatment” (1992: 455). This ‘new penology’ is a framework for analyzing the logic of risk, and understanding various techniques for identifying and managing ‘populations’ within criminal justice. This approach, they argue, is considered to be cost-effective, rational, and directed towards the future rather than the past (see also Feeley and Simon 1995).

Feeley and Simon maintain that the logic of risk has resulted in imprisonment being imposed primarily for the purposes of incapacitation – largely an instrumental function. The development and increase in the use of ‘intermediate’ sanctions in the 1990s, they argue, is an extension of the continuum of control. The most ‘risky’ and dangerous offenders are held in prison, and controls are extended beyond the prison, such as through the development of ‘intermediate’ sanctions like intensive probation, parole, community service, boot camps and others. ‘Intermediate’ sanctions provide low-cost surveillance for low-risk offenders (1992: 459; see also Feeley and Simon 1995).

In providing an analysis of the discourse of risk in the contemporary scene, they suggest that all punishments stem from this justification. However Feeley and Simon (1992, 1995) do not acknowledge that a particular sanction, such as the prison, might be seen as appropriate because it is perceived to effectively accomplish expressive or symbolic functions of punishment in addition to being instrumental for a minor violent offence, rather than being seen to accomplish the management of ‘high-risk’ offenders (see Shearing 1997; Hannah-Moffat 1999). As Garland (1990) and others (Shearing 1997; Hannah-Moffat 1999) suggest, it is important to highlight that Feeley and Simon (1992) suggest the concept of ‘the new penology’ as a framework or analytical tool – compared to a program – through which to understand developments in criminal justice (see page 460).
1997) have noted, punishment can symbolize condemnation of behaviour for instrumental ends.

Joel Feinberg (1970), for example, maintains that punishment involves pain, but also has an expressive function. He states that "punishment is a conventional device for the expression of attitudes of resentment and indignation, and of judgments of disapproval and reprobation, on the part either of the punishing authority himself or of those 'in whose name' the punishment is inflicted" (Feinberg 1970: 74). In fact Canadian case law suggests that imprisonment in particular is relied upon in order to accomplish the expressive purposes of sentencing — denunciation and general deterrence (R. v. G.M. (1992)). More recently in the context of conditional sentences, the courts have grappled with the question of whether sanctions within the community can accomplish the purpose of denunciation as effectively as imprisonment (in the traditional prison setting) (R. v. Proulx (S.C.C.) (2000); R. v. Wismayer (Ont. C.A.) (1997)).

It appears that the literature on these two paradigm shifts focus on broad developments within and in varying degrees beyond the criminal justice system. In the context of the criminal justice system, there is little discussion about other purposes — in addition to the single focus of either rehabilitation or managing risk — that are achieved by particular sanctions and how these connections might vary according to the nature of the behaviour being punished.

Variation in Punishment Across Time and Place

Most authors representing this fourth set of literature have focused on understanding the acceptability of certain punishments historically and across cultures.
While analysts have addressed the fact that different punishments have different functions over time, it appears that what is lacking is a framework for analyzing how a specific punishment (e.g., imprisonment or the fine) is seen as accomplishing specific purposes only with respect to certain offences. This set of literature comes closest to incorporating various dimensions of punishment -- including the sanction, purpose, and offence -- into analyses.

Some anthropological literature on punishment has revealed that traditional cultural values about community, interdependency, and social identity have been foundations that sustain informal social controls and restorative approaches to dealing with disputes in places such as Japan (Braithwaite 1989), Kilimanjaro (Moore 1986), and Morocco (Rosen 1989). For example, Rosen's analysis of justice in Islamic society reveals that cultural values of negotiation and interdependency among people are reinforced within the court process and system of 'punishment' in Morocco (1989). In the Western context, literature such as David Garland's *Punishment and Modern Society* (1990) suggests that by examining sensibilities about violence, civility and punishment, it possible to understand how different criminal justice punishments are seen as serving different purposes over time and place:

The ways in which we punish depend not just on political forces, economic interests, or even penological considerations but also on our conceptions of what is or is not culturally and emotionally acceptable. Penal policy decisions are always taken against a background of mores and sensibilities that, normally at least, will set limits to what will be tolerated by the public or implemented by the penal system's personnel. Such sensibilities force issues of 'propriety' on even the most immoral of governments, dictating what is and is not too shameful or offensive for serious consideration.

There is thus a whole range of possible punishments (tortures, maimings, stonings, public whippings, etc.) that are simply ruled out as 'unthinkable'
because they strike us as impossibly cruel and 'barbaric' -- as wholly out of keeping with the sensibilities of modern, civilized human beings. Such judgments, based on prevailing sensibilities, define the outer contours of possibility in the area of penal policy. Usually this boundary line has the unspoken, barely visible character of something that everyone takes for granted. It becomes visible, and obvious, only when an outrageous proposal crosses the line, or else when evidence from other times or other places shows how differently that line has been drawn elsewhere. It is therefore stating the obvious -- but also reminding us of something we can easily forget -- to say that punishments are, in part, determined by the specific structure of our sensibilities, and that these sensibilities are themselves subject to change and development (Garland 1991:142-43).

According to Garland, then, what might be considered as an acceptable intermediate sanction in one historical context, might be thought of as inappropriate at another point in time because of shifts in sensibilities about punishment. Beattie (1986) demonstrated how changing sensibilities about violence and public forms of punishment had an impact on the development and expansion of transportation as a 'secondary' punishment in 1718. In addition, Beattie highlights the emergence of transportation as a needed secondary punishment to fill the gap for less serious felonies but not serious enough for hanging. More broadly he states:

Explaining them, or rather explaining why alternatives [to hanging] were sought (by whom and for what purpose) and the success of some over others, is much more difficult because changes in punishment are almost certain not to arise from a simple, one-dimensional effect. The forms of punishment employed by a society at any moment are shaped by a variety of interests and intentions....Why one method of punishment loses favor over time and gives way to another is a complex question because penal methods evolve within a larger social and cultural context that in imperceptible ways alters the limits of what is acceptable in that society and what is not (Beattie 1986: 470).

Both Beattie and Garland argue that the acceptability of punishment should be understood within a broad social context, and beyond a perspective which focuses exclusively on instrumental criminal justice purposes (denunciation, deterrence,
rehabilitation, etc.), or solely on the severity of sanctions. Rather they suggest that punishment is multidimensional. Garland (1990) addresses the multidimensionality of punishment by using the example of the role of imprisonment in the twentieth century:

[crime] control — in the sense of reforming offenders and reducing crime rates — is certainly one of [the] objectives of [imprisonment], but by no means the only one....Most important, the prison provides a way of punishing people — of subjecting them to hard treatment, inflicting pain, doing them harm — that is largely compatible with modern sensibilities and conventional restraints on open, physical violence. In an era when corporal punishment has become uncivilized, and open violence unconscionable, the prison supplies a subtle, situational form of violence against the person that enables retribution to be inflicted in a way that is sufficiently discreet and 'deniable' to be culturally acceptable to most of the population. Despite occasional suggestions that imprisonment is becoming too lenient — a view that is rarely shared by informed sources — it is widely accepted that the prison succeeds very well in imposing real hardship, serious deprivation and personal suffering on most offenders who are sent there.

In terms of penological objectives then, the prison supports a range of them, and is "functional" or "successful" with respect to some, less so with respect to others....

Consequently...if one wishes to understand the prison as an institution — and the same arguments apply to the fine, probation, the death penalty, and the rest — it does little good to do so on a single plane or in relation to a single value. Instead, one must think of it as a complex institution and evaluate it accordingly, recognizing the range of its penal and social functions and the nature of its support (Garland 1990: 289-90).

More recently, Duff and Garland (1994) discuss sensibilities about 'intermediate' sanctions in particular, alluding to the importance of addressing the integration of purposes, sanctions, and offences. They state:

On both sides of the Atlantic governments have sought to persuade sentencers to reduce their use of imprisonment as a sanction. In the USA, this has been done by the use of statutory sentencing guide-lines...These reductionist efforts are frequently undercut by the law and order rhetoric which ministers deploy in public: but they also pose new kinds of problem for penal philosophy. What role should such sanctions play (and correctly, what role should imprisonment play) in the penal armoury? What goals or principles should guide their use? In what kinds of cases should they be used? Can we, or should we, try to apply them in accordance with the principle of proportionality — which would require us to rank these very different kinds of sanctions on a single scale of severity, in order
to ensure the severity of punishment is proportionate to the seriousness of the crime? (Duff and Garland 1994: 20).

Building on the perspectives of Garland, Beattie, and Duff, this thesis argues that the perceived appropriateness of interchangeable sanctions might be more challenging than first assumed, and suggests that model must include a neglected feature of the literature on sentencing and punishment -- the nature of the offence. In order to understand the acceptability of punishments, one has to look not only at the particular punishment and purpose it is supposed to serve, but the offence that the pain is being imposed for. Further dimensions of punishment need to be addressed-- such as characteristics of the offence and the offender -- in addition to the purposes and functions of individual punishments.

Data from one research study acts a springboard for this argument and attempts to build on past literature and approaches to understanding the perceived appropriateness of punishment. A relatively small sample of members of the public in Toronto\(^6\) was asked about the appropriateness of substituting a fine for a sentence of imprisonment across various offences (Marinos 1997; see also Doob and Marinos 1995). For a minor theft, members of the public saw imprisonment as significantly more effective as a form of denunciation, compared to fines. In fact while different sizes of fines and lengths of sentences of imprisonment were presented to respondents, the type of the penalty -- fine or prison, independent of size -- had a significant impact on public perceptions of the denunciatory value (Tables 1 & 2, Marinos 1997). These findings provided a context for

---

\(^6\) One hundred and twenty members of the public agreed to respond to the questionnaire about their views on interchangeability of fines with sentences of imprisonment. The sample was composed of a relatively equal number of adult males (46.7%) and females (53.3%), ranging from ages 18 to over 60 years old. A question pertaining to family income indicated that the sample was made up of people from a variety of economic backgrounds.
understanding public acceptability of interchangeability of fines and sentences of imprisonment across various offences.

Respondents were also provided with a range of twelve different offences (five property, four violent, and three other offences), and asked to comment on the appropriateness of the use of a fine -- of any size -- instead of a sentence of imprisonment for each offence. The data revealed that the acceptability of fines varied according the nature of the offence. Presumably offences which were considered to be violent, sexual, or dangerous were not widely seen as being suitable for a fine of any size. In contrast, respondents supported the use of fines as substitutes for a sentence of imprisonment for offences involving theft or damage of property, and possession of marijuana.

What is significant about this study is the finding of the perception that fines do not denounce crime as effectively as imprisonment for particular violent offences. Therefore this is one piece of evidence to suggest that different punishments are seen to have different functions according to the nature of the offence. This prevents the public from seeing them as being equivalent to imprisonment, and may have the result that fines are not as appropriate whenever denunciatory statements are needed. It appears that for offences involving violence and harm to the body, compared to property crimes, members of the public believe that the expression of condemnation or denunciation is an important purpose, and is viewed as most effectively accomplished through imprisonment. Finally it suggests that there are qualitative in addition to quantitative dimensions to punishment.
Moving Beyond the Dimension of Severity:

Multidimensionality and the Role of the Nature of the Offence

Past research evidence on the public's perceived appropriateness of fines as interchangeable sanctions with imprisonment suggests that we must develop more complex models of sentencing than a focus on the seriousness of the offence and the severity of sanctions to accomplish a range of purposes of sentencing. One example is the U.S. sentencing guidelines in some states such as Minnesota, Washington state, and others (see Morris and Tonry 1990; Tonry and Hatlestad 1997; Miller 1999). In the case of sentencing guidelines in some states, it is rather common that offences are put on a single scale of seriousness, and that the decision is on “prison” or not, rather than on which penalty is appropriate. Proportionality in sentencing is another example of a sentencing model that focuses primarily on severity. As mentioned earlier, while Canadian sentencing legislation includes the principle of proportionality in the context of the purposes of sentencing, there is no explicit link made between the ability of a punishment to accomplish a particular purpose and to be proportional to the offence and harm done.

The past research findings, and the literature by Garland (1990) and others (Beattie 1986; Duff and Garland 1994), however, suggest that sentencing models should include some consideration of the nature of the offence in addition to its seriousness to determine the appropriate purpose to be achieved by a particular punishment. Thus far it appears that Canada's current model to reduce the use of imprisonment through

---

7 The twelve offences included: “steal VCR”, “steal ring”, “possession of stolen goods”, break windows”, “fraud, stolen credit card”, “assault, broken bottle”, “robbery”, “minor sexual assault”, “sexual assault”, “possession marijuana”, “selling cocaine”, and “impaired driving” (see Table 3, Marinos 1997 at 39).
‘intermediate’ sanctions has not addressed the possible connections among sanctions, purposes, and the quantity and qualities of the offence and offender.

Sanctions such as fines, community service orders, electronic monitoring, intensive supervision and others have generally been referred to as ‘alternatives to imprisonment’ or ‘community corrections’. Implicit in these terms is the categorization of non-prison punishments as the ‘other’ of imprisonment. Since the late 1980s, however, these punishments are more likely to be referred to as ‘intermediate’ within both the academic literature and policy documents (Daubney 1988; Sentencing Reform Team 1991; Morris and Tonry 1990).

The change in terminology is significant. The underlying assumption is that in order to shift attention away from imprisonment and build a focus on other sanctions, it is important to constitute individual ‘intermediate’ punishments as being important in and of themselves. The concept of ‘intermediate’ sanctions, then, attempts to accomplish this shift from conceptualizing the prison and the ‘other’, to emphasizing individual attention to a range of individual sanctions between prison and probation (see Morris and Tonry 1990). Thus intermediacy in punishment is not an issue of semantics, but rather represents a significant shift in conceptualizing punishment.

In 1996, changes were made to the Canadian Criminal Code with respect to fines and community service orders, and a new sanction, the conditional sentence of imprisonment, was introduced in the package of reforms. In total, the changes made to these punishments were part of a larger shift towards creating and increasing the use of ‘intermediate’ sanctions in Canada, with the hope of replacing and therefore decreasing the use of imprisonment. In terms of fines, judges must inquire into the offender’s means
to pay a fine before it is imposed (section 734(2)). The option of imposing up to 240 hours of community service orders over 18 months as a condition of probation is explicitly listed (section 732.1(3)(f)). In the past, community service orders were handed down as part of probation, but community service was not legislated as a specific optional condition within the Code. Lastly the conditional sentence of imprisonment -- to be served within the community -- was developed to replace, or be interchangeable with traditional sentences of imprisonment of up to two years (section 742). It is assumed, then, that punishments can be made equivalent and therefore interchangeable by adjusting the severity of the punishment.

It may be possible to create equivalencies among different punishments in terms of severity, but it appears from past research that members of the public see different qualities among sanctions (Marinos 1997; Doob and Marinos 1995). While the changes have been made to fines, community service orders, and conditional sentences, the problem remains that there is no specific guidance for judges or members of the public on how particular ‘intermediate’ sanctions might be more capable than others in accomplishing a particular purpose of sentencing for a particular offence and offender (youth or adult) -- beyond an assumption about the severity of sanctions.

Most recently, the proposed Youth Criminal Justice Act (Bill C-3, Second Session, Thirty-sixth Parliament) addresses the complexities of punishment one tiny step further. There is an explicit differentiation among violent and non-violent cases in section 38. One of the conditions that allow for a prison sentence is violence. Second, there is an acknowledgement that “within a level of severity” different sanctions can accomplish different goals (e.g., rehabilitation and reintegration). Thus in this most recently proposed
legislation, purposes are associated with punishment to some extent, and the type of offence is associated to some extent.

This thesis examines public and judicial sensibilities about the appropriateness of interchangeable sanctions with sentences of imprisonment. I address some of the perceived connections made among the functions of punishment, the purposes of sentencing, and various sanctions in the context of the offence being punished. This last dimension of punishment -- the nature of the offence -- has not been highlighted as an important feature when attempting to understand the perceived appropriateness of punishment. *This thesis serves to address the following question: What are some of the functions that punishments are seen as accomplishing by members of the public and judges in the context of the purposes of sentencing, the nature of offence, and the characteristics of the offender?*

The next four Chapters address public and judicial sensibilities about interchangeability. These two 'groups' were selected because they are both essential to understanding the culture of punishment. While members of the public believe that some punishments are inappropriate compared to others in different contexts, judges are part of the public and at the same time are responding to concerns from public and judicial cultures. The thesis question is addressed by focusing on public sensibilities towards fines and community service orders (Chapter 2), as well as conditional sentences of imprisonment (Chapter 3). Sentencing practices are examined in Chapter 4, using youth court dispositions as a site. Lastly Chapter 5 examines judicial sensibilities -- at both the trial and appeal court levels -- towards conditional sentences of imprisonment. An examination of public and judicial sensibilities towards interchangeability will not only
build on past work about understanding the perceived appropriateness of punishment, but also add to multidimensional perspectives on punishment.
Chapter 2

Public Sensibilities About Interchangeability:
Fines and Community Services Orders

Introduction

The underlying assumption of interchangeability is that sanctions can be placed along a continuum of punishments, with probation on one end and prison on the other end (Morris and Tonry 1990: 40). Authors Morris and Tonry (1990) argue that ‘intermediate’ sanctions -- such as fines and community service orders -- can be made equivalent to imprisonment and thus interchangeable across a range of offences. They suggest, however, that ‘intermediate’ sanctions could be substituted for imprisonment only for sentences of up to two years. In the United States, up to two years imprisonment is justified for “nonviolent or not-very-violent stranger crime” (1990: 79).

Morris and Tonry (1990) devoted special attention to two sanctions in particular -- fines and community service orders -- as the most promising ‘intermediate’ sanctions for interchangeability with sentences of imprisonment (see similar discussions by Byrne, Lurigio, and Peterselia 1992; Smykla and Selke 1995; Tonry and Hamilton 1995). As mentioned in Chapter 1, there have been legal changes made to fines and community service orders in the Code which make it easier for the courts to impose punishments other than, or as substitutes to imprisonment. However, sentencing legislation for adults and youth do not make explicit links among sanctions, purposes of sentencing, and offences.

1 Morris and Tonry (1990) explain probation as being the least severe sanction in this way: “Nor shall we deal at any length with normal or ordinary probation, which for state and local crime in many cities has degenerated into ineffectiveness under the pressure of excessive caseloads and inadequate resources. To be sure, there may well be many times when nominal probation, giving the appearance but not the reality of punishment, is exactly what’s wanted” (Morris and Tonry 1990: 6).
While finding alternatives to imprisonment are essential, there are limitations to how interchangeability has been conceptualized. First, there are a number of assumptions about the qualities and characteristics of sanctions and what they entail. It is assumed that sanctions should be assessed solely in terms of their size or severity. In fact research on perceptions of ‘intermediate’ sanctions has generally focused on comparing and ranking sanctions by their severity (Tremblay 1988; Peterselia and Deschenes 1994; Harlow, Darley and Robinson 1995). Since the size or severity is the only factor to consider, it is assumed that sanctions are flexible enough to be made equivalent and substituted. Second, it is assumed that interchangeability can be accomplished across a wide range of offences (property and violent), and for a range of sentencing purposes.

It is precisely the focus on conceptualizing ‘intermediate’ sanctions and interchangeability as solely quantitative issues -- and as a single dimension of severity -- which is most problematic. As mentioned in Chapter 1, a research study on public perceptions of interchangeability of fines for sentences of imprisonment (Doob and Marinos 1995; Marinos 1997) revealed that imprisonment was relied upon by members of the public in order to accomplish denunciation for violent offences (within the limits of this experiment). This purpose appeared to be seen as an important objective of sentencing in offences involving violence. Garland (1990), for instance, argues that there are various dimensions to punishment -- moral, political, economic, and cultural -- and to evaluate punishment in merely instrumental terms is both misguided and unproductive.

Understanding that there are qualitative as well as quantitative dimensions to punishment raises the question of acceptability. Thinking about interchangeability on a

\[^2\text{In fact, Michael Tonry suggests, in his report on intermediate sanctions for the National Institute of Justice, that intermediate sanctions be included into sentencing guidelines (1997: 29).}\]
single continuum — that only severity of sanctions is important — would lead one to believe that the public would be very understanding of substituting a number of different ‘intermediate’ sanctions for sentences of imprisonment. But if one assumes that sanctions, and indeed punishment include dimensions other than severity, then it is possible to think about various qualitative factors which might be thought of as inappropriate by the public. For instance, are sanctions perceived to be equally interchangeable for different offences? Are sanctions equally interchangeable for adult and youth? When people are willing to substitute a fine or community service order for a sentence of imprisonment, does the size of the penalty vary across different offences? What other beliefs about sentencing offenders — such as the purposes of different punishments — are related to the public’s support for interchangeability?

In this Chapter, I challenge the simplistic notion that perceptions of punishment can be understood as a single dimension of severity. This is examined through public sensibilities toward substituting fines and community service orders for sentences of imprisonment. I argue that it is inaccurate to assume that sanctions are easily interchangeable. They cannot be placed along a single continuum where a certain amount of one ‘intermediate’ sanction can be substituted for a sentence of imprisonment by simply focusing on severity. One also cannot assume that if a sanction can be substituted for one offence, it can be substituted for others. In exploring qualitative in addition to the quantitative dimensions, this Chapter will demonstrate that members of the public have limits about the acceptability of interchangeability based upon the nature of the offence, the nature of the ‘intermediate’ sanction, and the purposes of punishment.

The findings presented in this Chapter were derived from a survey of 1,006
households in Ontario. Respondents were randomly sampled adults living in Ontario who were interviewed by telephone. They were chosen by a survey company using a random digit dialing technique. Five hundred (500) people responded to the questionnaire on adult offenders, and a separate sample of five hundred and six (506) Ontario residents responded to questions on young offenders. On the whole, the questionnaire attempted to understand the public's sensibilities about a range of 'intermediate' sanctions, such as fines, community service orders, and conditional sentences of imprisonment (presented in Chapter 3), in comparison to sentences of imprisonment. Since the multidimensionality of punishment is the focus of these analyses, it was important to explore public sensibilities relating to both adult and young offenders when possible.

---

3 Goldfarb and Associates carried out the telephone surveys.
4 The term “sensibilities” refers to the feelings, values, and ways of thinking which are associated with, and support attitudes. Exploring “sensibilities” is useful in the context of attitudes toward interchangeability because it exposes ways of thinking, feeling, and emotions about different sanctions and what it means to punish. See David Garland, 1990, for a broader discussion of sensibilities and its relationship to “culture”.
Findings

_are sanctions viewed as equally interchangeable for different offences?_

The first way in which the issue of interchangeability was explored was to see whether there was variation in respondents’ willingness to substitute a fine and a community service order for a sentence of imprisonment. A question in the survey asked respondents: “A judge is thinking of putting an adult/young person) in prison for 30 days for [stealing CDs worth about $200/assault/sexual assault]. Instead of a prison sentence, do you think that a fine of some size could be appropriate? Respondents were asked about only one of three offences, and therefore approximately one-third of the respondents were asked about each offence. If they supported the substitution of a fine and/or a community service order for the thirty day prison sentence, then they were asked about the amount of a fine and the number of hours which could be appropriate. Tables 2.1 and 2.2 present findings of public support for substituting a fine or community service order for a thirty day prison sentence for three different offences for an adult offender.
Table 2.1
Percent of Respondents Willing to Substitute a Fine or a Community Service Order For a Thirty Day Prison Sentence For Three Different Offences For an Adult Offender

<table>
<thead>
<tr>
<th>Support For Interchangeability</th>
<th>Offence</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Steal CDs</td>
<td>Assault</td>
<td>Sexual Assault</td>
</tr>
<tr>
<td>Substitution of Fine for Prison</td>
<td>68.5%</td>
<td>40.9%</td>
<td>46.3%</td>
</tr>
<tr>
<td>Substitution of CSO for Prison</td>
<td>80.0%</td>
<td>71.5%</td>
<td>67.2%</td>
</tr>
</tbody>
</table>

Table 2.1 demonstrates (in the adult survey) that respondents are much more likely to support interchangeability of a fine or community service order for a 30 day prison sentence for the property offence of theft of CDs (described as “worth about $200”), than for an assault (described as “assault”) or a sexual assault (described as a “relatively minor sexual assault for example, where a man touched a woman’s breasts without her consent”). Thus the findings presented in Table 2.1 reveal that support for substituting fines or community service orders for sentences of imprisonment varies according to the nature of the offence — a qualitative dimension of punishment. In this way, sanctions are not equally interchangeable for different offences, even though the starting points in all cases was a thirty day prison sentence.
Are sanctions viewed as qualitatively different from each other?

What can also be seen from these findings (Table 2.1) is that there is greater support for community service orders compared to fines for each of the three offences. If sanctions were viewed as differing only on a single dimension of severity, however, then one would expect that respondents would not be making distinctions among the acceptability of different ‘intermediate’ sanctions as alternatives to prison. While there were significant correlations between the probability of both ‘intermediate’ sanctions being seen as interchangeable with a 30 day sentence of imprisonment across the three offences, community service was significantly more likely to be supported as a substitute for prison than were fines (theft: $\chi^2 = 4.89$; df= 1; p<.05; assault: $\chi^2 = 33.35$; df= 1; p<.05; sexual assault: $\chi^2 = 62.02$; df=1; p<.05). This finding suggests that members of the public perceive there to be differences between the two sanctions in terms of characteristics and qualities, and this affects the nature and extent of support for substitution for a sentence of imprisonment. In this way, sanctions are not seen as equally interchangeable because of their very nature, characteristics, and qualities.
Are sanctions seen as equally interchangeable for adult and youth offenders?

The findings presented thus far have related solely to the adult survey. As mentioned earlier, respondents were asked to answer either a set of questions focusing largely on adult offending or on youth offending. One might expect that if members of the public focus only on the severity of a sanction to evaluate its appropriateness, then it is possible that perceptions about adult sentencing and youth sentencing would not be different, regardless of the nature of the offence. On the other hand, if a different pattern of acceptability of interchangeability resulted -- for adults compared to youth -- then it would support the view that the offending population (within the limits of this experiment) is an important dimension. Thus Table 2.2 presents results of the public's willingness to substitute a fine or community service order for a thirty day prison sentence for the offences of theft of CD's, assault, and sexual assault involving a young offender.
Table 2.2

Percent of Respondents Willing to Substitute a Fine or Community Service Order For Three Different Offences for a Young Offender

<table>
<thead>
<tr>
<th>Offence</th>
<th>Support For Interchangeability</th>
<th>Steal CDs</th>
<th>Assault</th>
<th>Sexual Assault</th>
<th>Significance of Difference Across Offences</th>
</tr>
</thead>
<tbody>
<tr>
<td>Substitution of Fine for Prison</td>
<td>61.1%</td>
<td>25.1%</td>
<td>43.8%</td>
<td>$\chi^2 = 43.11$, df = 2, p&lt;0.001</td>
<td></td>
</tr>
<tr>
<td>Substitution of CSO for Prison</td>
<td>84.6%</td>
<td>77.5%</td>
<td>75.9%</td>
<td>$\chi^2 = 4.56$, df = 2, Not significant</td>
<td></td>
</tr>
</tbody>
</table>

Table 2.2 demonstrates that respondents were significantly more likely to support the substitution of a fine and community service order for a 30 day prison sentence for the offence of theft, compared to assault or sexual assault, for a young offender. This finding is consistent with results presented in Table 2.1 where the nature of the offence -- property compared to violence -- was an important factor in respondents' acceptability of interchangeability for an adult offender. The next question to address is whether interchangeability is seen as equally acceptable for different offender populations -- adult or youth. The findings are presented in 2.3 and 2.4.

---

5 This finding was similar for a community service order but was not statistically significant ($\chi^2 = 4.56$, df=2; not significant).
Table 2.3
Percent of Respondents Willing to Substitute a Fine for the Offences of Stealing Cds, Assault, Sexual Assault, and Pooled Across Three Offences, for Adult and Young Offenders

| Willingness to Substitute Fines For . . . | Offence | | | | | Significance of Difference Across Offences |
|------------------------------------------|---------|---------|---------|----------------------------------------|
| Steal CDs | Assault | Sexual Assault | |
| **Adult Offender** | 68.5% (98) | 40.9% (67) | 46.3% (82) | 51.0% (247) |
| **Young Offender** | 61.1% (110) | 25.1% (43) | 43.8% (64) | 43.7% (217) |
| **Significance of Difference** | $\chi^2 = 1.604$, df = 1, not significant | $\chi^2 = 8.665$, df = 1, p<.01 | $\chi^2 = .113$, df = 1, not significant | $\chi^2 = 5.345$, df = 1, p<.05 |

Table 2.4
Percent of Respondents Willing to Substitute a Community Service Order for the Offences of Stealing Cds, Assault, Sexual Assault, and Pooled Across Three Offences, for Adult and Young Offenders

| Willingness to Substitute CSOs For . . . | Offence | | | | | Significance of Difference Across Offences |
|----------------------------------------|---------|---------|---------|----------------------------------------|
| Steal CDs | Assault | Sexual Assault | |
| **Adult Offender** | 80.0% (116) | 71.5% (118) | 67.2% (121) | 72.4% (355) |
| **Young Offender** | 84.6% (154) | 77.5% (134) | 75.9% (110) | 79.6% (398) |
| **Significance of Difference** | $\chi^2 = 0.895$, df = 1, not significant | $\chi^2 = 1.273$, df = 1, not significant | $\chi^2 = 2.510$, df = 1, not significant | $\chi^2 = 6.562$, df = 1, p<.01 |
Table 2.3 reveals that in examining the three offences, respondents are significantly less supportive of the substitution of fines for assault for youth, compared to adults. Overall (across pooled offences) fines are viewed as significantly less acceptable substitutes for prison for young offenders (43.7%), compared to adults (51.0%). Table 2.4, on the other hand, exhibits a reverse pattern: respondents view community service orders, across the three offences, as significantly less acceptable and substitutable for adult offenders (72.4%) compared to young offenders (79.6%). While the magnitude of these percentages for fines and community services are not large, they are nonetheless (statistically) significant. Looking at Tables 2.3 and 2.4 together, it is evident that respondents are significantly more likely to support the substitution of fines for adult offenders, and community service orders for young offenders. Thus consistent with the hypothesis, 'intermediate' sanctions such as fines and community service orders are not viewed as equally interchangeable across different offender populations -- as an adult or youth -- in addition to different offences. Clearly these findings also reinforce earlier results that the public perceive there to be qualitative differences between the two sanctions.

The Quantitative Dimension of Punishment:

When people are willing to substitute, do the size of the fine and hours of community service order vary according to the nature of the offence?

It also was necessary to explore the quantitative dimension of punishment, in addition to qualitative factors. Not only were respondents asked about substituting a fine or community service order for a thirty day prison sentence, but the question included a
second part: if the respondent believed that a fine or community service order was "substitutable", then they were asked to suggest the appropriate size/amount of fine and hours of community service. Tables 2.5 through 2.8 present the findings of the quantitative dimension of interchangeability.\(^6\)

### Table 2.5

**Amount of Fine Chosen as Substitute for a 30 Day Sentence of Imprisonment for Theft, Assault, and Sexual Assault Committed by Adult Offender (by Those Willing to Substitute)**

<table>
<thead>
<tr>
<th>Offence (Adult Offender)</th>
<th>&quot;Low&quot; Fine $1-$450</th>
<th>&quot;Medium&quot; Fine $500-$900</th>
<th>&quot;High&quot; Fine $1000-$30,000</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Steal CDs</td>
<td>44.8% (43)</td>
<td>35.4% (34)</td>
<td>19.8% (19)</td>
<td>100% (96)</td>
</tr>
<tr>
<td>Assault</td>
<td>13.1% (8)</td>
<td>19.7% (12)</td>
<td>67.2% (41)</td>
<td>100% (61)</td>
</tr>
<tr>
<td>Sexual Assault</td>
<td>32.1% (25)</td>
<td>25.6% (20)</td>
<td>42.3% (33)</td>
<td>100% (78)</td>
</tr>
</tbody>
</table>

\(\chi^2 = 36.53, \, df=4, \, p<.001\)

\(^6\) The amount of the fine and the hours of the CSO proposed by those respondents willing to substitute were broken down into three categories according to the median. The mean scores were not the best measures because the number of CSO hours were specified as 240. Thus the average would most likely cluster around this amount.
Table 2.6

Number of Hours of Community Service Order Chosen as Substitute for a 30 Day Sentence for Theft, Assault, and Sexual Assault Committed by Adult Offender (by Those Willing to Substitute)

<table>
<thead>
<tr>
<th>Offence (Adult Offender)</th>
<th>&quot;Low&quot; Hours CSO 5-175</th>
<th>&quot;Medium&quot; Hours CSO 180-200</th>
<th>&quot;High&quot; Hours CSO 240-5000</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Steal CDs</td>
<td>32.2% (37)</td>
<td>21.7% (25)</td>
<td>46.1% (53)</td>
<td>100%</td>
</tr>
<tr>
<td>Assault</td>
<td>28.3% (32)</td>
<td>8.8% (10)</td>
<td>62.8% (71)</td>
<td>100%</td>
</tr>
<tr>
<td>Sexual Assault</td>
<td>39.2% (47)</td>
<td>16.7% (20)</td>
<td>44.2% (53)</td>
<td>100%</td>
</tr>
</tbody>
</table>

$\chi^2 = 12.96$, df=4, p<.01

Table 2.7

Amount of Fine Chosen as Substitute for a 30 Day Prison Sentence for Theft, Assault, and Sexual Assault Committed by Young Offender (by Those Willing to Substitute)

<table>
<thead>
<tr>
<th>Offence (Young Offender)</th>
<th>&quot;Low&quot; Fine $25-$200</th>
<th>&quot;Medium&quot; Fine $225-$500</th>
<th>&quot;High&quot; Fine $600-$25,000</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Steal CDs</td>
<td>28.0% (30)</td>
<td>47.7% (51)</td>
<td>24.3% (26)</td>
<td>100%</td>
</tr>
<tr>
<td>Assault</td>
<td>27.0% (10)</td>
<td>29.7% (11)</td>
<td>43.2% (16)</td>
<td>100%</td>
</tr>
<tr>
<td>Sexual Assault</td>
<td>56.9% (33)</td>
<td>22.4% (13)</td>
<td>20.7% (12)</td>
<td>100%</td>
</tr>
</tbody>
</table>

$\chi^2 = 21.63$, df=4, p<.01
Table 2.8

Number of Hours of Community Service Chosen as Substitute for a 30 Day Prison Sentence for Theft, Assault, and Sexual Assault Committed by Young Offender (by Those Willing to Substitute)

<table>
<thead>
<tr>
<th>Offence (Young Offender)</th>
<th>&quot;Low&quot; Hours CSO 2-125</th>
<th>&quot;Medium&quot; Hours CSO 150-220</th>
<th>&quot;High&quot; Hours CSO 240-2,000</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Steal CDs</td>
<td>26.7% (40)</td>
<td>28.0% (42)</td>
<td>45.3% (68)</td>
<td>100% (150)</td>
</tr>
<tr>
<td>Assault</td>
<td>22.7% (29)</td>
<td>22.7% (29)</td>
<td>54.7% (70)</td>
<td>100% (128)</td>
</tr>
<tr>
<td>Sexual Assault</td>
<td>45.8% (49)</td>
<td>16.8% (18)</td>
<td>37.4% (40)</td>
<td>100% (107)</td>
</tr>
</tbody>
</table>

χ² = 18.66, df=4, p<.01

The findings presented in Table 2.5 demonstrate that when people are willing to substitute a fine for prison in the case of an adult offender, the amount of the fine is determined by the nature or apparent seriousness of the offence. On the whole the four tables presented above (Tables 2.5 through 2.8) reveal consistent findings related to the offence of assault. In examining these findings, it is clear that of those willing to substitute, there is a higher proportion of respondents who suggest a "high" amount or hours for assault compared to the other two offences. It is likely that respondents perceived this offence to be the most serious offence compared to the theft and the (minor) sexual assault. In addition, this conclusion is reinforced in Tables 2.1 and 2.2 by the lower level of support for the substitution of a fine for assault, compared to the other two offences.
Is willingness to support interchangeability, then, related to other qualitative dimensions of punishment?

The following section presents findings of those willing to substitute fines or community service orders -- compared to non-supporters -- for a thirty day prison sentence across the three offences (offences pooled). The purpose is to understand whether supporters of interchangeability might have other beliefs about punishment in Canada which might relate to their preference for substitution. These broader relationships provide an understanding of who and why interchangeability is supported or opposed, as well as reveal other dimensions of punishment.

The first relationship explored was whether support for interchangeability was related to confidence that sanctions in the community -- in particular a community service order -- are carried out and completed. If respondents have confidence in community service orders, then they are likely to respond positively to interchangeability. Therefore respondents were asked whether they think that “such orders [community service orders -- where the offender is ordered to do a certain number of hours of unpaid work in the community such as working for a church, animal shelter, or food bank or removing graffiti in public places] are fully completed “always”, “most of the time”, “about half of the time”, “rarely”, or “almost never”. It is expected that those willing to support substitution of fines and community service orders would be more likely to believe that community service orders are carried out or completed always or most of the time, compared to non-supporters of interchangeability. The findings are presented in Tables 2.9 and 2.10.
Table 2.9

Respondents' Willingness to Support Interchangeability of a Community Service Order for a Sentence of Imprisonment and Their Perceptions of Completion Rates for Community Service Orders for Adult Offenders (Pooled Offences)

<table>
<thead>
<tr>
<th>Support &amp; Opposition Toward Interchangeability</th>
<th>Always/Most of the Time</th>
<th>About Half the Time</th>
<th>Rarely/Almost Never</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Support for Substitution of CSO</td>
<td>36.3% (124)</td>
<td>38.9% (133)</td>
<td>24.9% (85)</td>
<td>100% (342)</td>
</tr>
<tr>
<td>Opposition to Substitution of CSO</td>
<td>33.6% (44)</td>
<td>29.8% (39)</td>
<td>36.6% (48)</td>
<td>100% (131)</td>
</tr>
</tbody>
</table>

$\chi^2 = 7.036, df= 2, p<.01$

Table 2.10

Respondents' Willingness to Support Interchangeability of a Community Service Order for a Sentence of Imprisonment and Their Perceptions of Completion Rates for Community Service Orders for Young Offenders (Pooled Offences)

<table>
<thead>
<tr>
<th>Support &amp; Opposition Toward Interchangeability</th>
<th>Always/Most of the Time</th>
<th>About Half the Time</th>
<th>Rarely/Almost Never</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Support for Substitution of CSO</td>
<td>36.5% (136)</td>
<td>42.4% (158)</td>
<td>21.2% (79)</td>
<td>100% (373)</td>
</tr>
<tr>
<td>Opposition to Substitution of CSO</td>
<td>38.1% (37)</td>
<td>25.8% (25)</td>
<td>36.1% (35)</td>
<td>100% (131)</td>
</tr>
</tbody>
</table>

$\chi^2 = 12.55, df= 2, p<0.01$

As expected, the findings presented in Tables 2.9 and 2.10 demonstrate that
support for substituting community service orders for both adult and young offenders across the three offences is related to perceptions of a relatively high level of community service completion. Those willing to support community service orders were significantly more likely to believe that the orders are completed “always” or “most of the time” compared to non-supporters. However, the same pattern did not emerge for support for interchangeability of fines. Since the question addressed completion of community service orders specifically, it is not surprising that the same relationship was not found for supporters of fines for substitution.

It is also possible that those who oppose interchangeability of fines and community service orders have a belief in the effectiveness of prisons to accomplish something, which impede them from supporting non-prison sanctions. One way to understand whether respondents have a belief in prisons is to ask questions about how they would like the government to invest money in criminal justice. This issue was measured by two indicators. For both the adult and youth surveys, respondents were asked two questions:

“Ontario’s prison are overcrowded. Two solutions that have been proposed are the following: (a) build more prisons and (b) sentence more offenders to alternatives to imprisonment such as probation, restitution, fines and community service orders (where the person does a specified number hours of work for an organization like a church or food bank). How do you think the government should spend its money? Build more prisons or sentence more offenders to alternatives to imprisonment (probation, restitution, fines, community service orders).”

“As I just pointed out, Ontario’s prisons are full. If the government were to have a sum of money to spend on crime, would you suggest that they spend it on more prisons or that they spend it on programs to prevent crime? More prisons or crime prevention.”

---

7 adult survey: fine by cso orders completed: chi-square, 1.164, df= 2, not significant; youth survey: fine by cso orders completed, chi-square, 1.21; df= 2, not significant.
The hypothesis is that those who support interchangeability of fines and community service orders are more likely to support the government spending public money on sentencing to alternatives to imprisonment as well as crime prevention strategies, instead of building more prisons. If there is a relationship between interchangeability and these broader views about spending money for criminal justice, then it will explain why some respondents support and oppose substitution. Each question is explored separately, for adult and young offenders (offences pooled), presented in Tables 2.11 through 2.14.

Table 2.11

Willingness to Support Interchangeability of a Community Service Order for a Sentence of Imprisonment and Perceptions of Government Investing in Building More Prisons or Alternatives to Imprisonment for Adult Offenders (across pooled Offences)

<table>
<thead>
<tr>
<th>Support &amp; Opposition Toward Interchangeability</th>
<th>Building More Prisons</th>
<th>Sentencing to Alternatives to Imprisonment</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Support for Substitution of CSO</td>
<td>30.1% (101)</td>
<td>69.9% (235)</td>
<td>100% (336)</td>
</tr>
<tr>
<td>Opposition to Substitution of CSO</td>
<td>46.9% (60)</td>
<td>53.1% (68)</td>
<td>100% (128)</td>
</tr>
</tbody>
</table>

$\chi^2 = 10.84$, df= 1, p<.001
Table 2.12
Willingness to Support Interchangeability of a Community Service Order for a Sentence of Imprisonment and Perceptions of Government Investing in Building More Prisons or Alternatives to Imprisonment for Young Offenders (Pooled Offences)

<table>
<thead>
<tr>
<th>Support &amp; Opposition Toward Interchangeability</th>
<th>Building More Prisons</th>
<th>Sentencing to Alternatives to Imprisonment</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Support for Substitution of CSO</td>
<td>16.1% (61)</td>
<td>83.9% (318)</td>
<td>100% (379)</td>
</tr>
<tr>
<td>Opposition to Substitution of CSO</td>
<td>43.2% (41)</td>
<td>56.8% (54)</td>
<td>100% (95)</td>
</tr>
</tbody>
</table>

χ² = 31.36, df= 1, p<.0001

Table 2.13
Willingness to Support Interchangeability of a Community Service Order for a Sentence of Imprisonment and Perceptions of Government Investing in Building More Prisons or Crime Prevention for Adult Offenders (Pooled Offences)

<table>
<thead>
<tr>
<th>Support &amp; Opposition Toward Interchangeability</th>
<th>Building More Prisons</th>
<th>Crime Prevention</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Support for Substitution of CSO</td>
<td>10.3% (36)</td>
<td>89.7% (313)</td>
<td>100% (349)</td>
</tr>
<tr>
<td>Opposition to Substitution of CSO</td>
<td>24.8% (32)</td>
<td>75.2% (97)</td>
<td>100% (129)</td>
</tr>
</tbody>
</table>

χ² = 15.04, df= 1, p<.0001
Table 2.14

Willingness to Support Interchangeability of a Community Service Order for a Sentence of Imprisonment and Perceptions on Government Investing in Building More Prisons or Crime Prevention for Young Offenders (Pooled Offences)

<table>
<thead>
<tr>
<th>Support &amp; Opposition Toward Interchangeability</th>
<th>Building More Prisons</th>
<th>Crime Prevention</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Support for Substitution of CSO</td>
<td>7.4% (29)</td>
<td>92.6% (363)</td>
<td>100% (392)</td>
</tr>
<tr>
<td>Opposition to Substitution of CSO</td>
<td>26.6% (25)</td>
<td>73.4% (69)</td>
<td>100% (94)</td>
</tr>
</tbody>
</table>

χ² = 26.38, df= 1, p<.0001

Consistent with expected results about the contrast between community service orders and fines, Tables 2.11 to 2.14 reveal that support for substitution of community service orders (only) was significantly related to investing in alternatives (Tables 2.11 and 2.12) and crime prevention strategies (Tables 2.13 and 2.14) for both adult and young offenders. Despite the fact that a range of individual sanctions (probation, restitution, fines, community service orders), including fines, were described to the respondent as part of “alternatives to imprisonment”, this pattern of results did not emerge for those willing to substitute a fine. Clearly respondents view community service orders and fines as qualitatively different.

Another factor that was explored was whether support for interchangeability of

---

8 adults: fine by prisons or alternatives -- chi-square, 0.619, df= 1, not significant; youth: fine by prisons or alternatives -- chi-square, 0.680, df= 1, not significant; adults: fine by prisons or crime prevention -- chi-square, 0.56, df= 1, not significant; youth: fine by prisons or crime prevention -- chi-square, 0.097, df= 1,
either a fine or community service order might be related to a belief that sentencing in general in Canada is too lenient. If those who oppose interchangeability also believe that sentencing is too lenient, then it would explain their unwillingness to substitute. This explanation would also be consistent with the theory that sanctions are understood as only differing by their severity. In order to explore this possibility, respondents were asked: "thinking about adult courts (or courts that deal with young offenders) in general, would you say that sentences handed down by the courts are too severe, about right, or not severe enough?"

For both adult and young offenders and across the three offences, support for interchangeability of fines or community service orders did not relate to views on the severity of sentencing. There was a tendency for supporters of interchangeability of both fines and community service to believe that sentences are too severe or about right, compared to non-supporters. However these findings were not statistically significant, and therefore cannot account for explaining support. Thus a quantitative assumption about the severity or leniency of sentencing in Canada does not help explain support or opposition to interchangeability.

Thus far, opposition to interchangeability is not explained by perceptions about the leniency of sentencing, yet is related to support for building more prisons. However it is essential to understand what aspect of prisons makes it an attractive choice for the public. It is possible that those who oppose interchangeability -- and thus support prisons

---

9 The only statistically significant finding is for youth: cso by youth court sentences -- chi-square, 4.11, df=1; p<.05.

Adult survey: fine by adult court sentences -- chi-square, 2.91, df=1; not significant, 0.08; cso by adult court sentences -- chi-square, 2.43, df=1, not significant, 0.118.

Youth survey: fine by youth court sentences -- chi-square, 0.44, df=1; not significant, 0.505.
-- see denunciation as an important purpose of sentencing. I argue that part of understanding the perceived appropriateness of punishment is that different sanctions are appropriate for different purposes, as well as for different offences and offenders. As mentioned in Chapter 1, prison is often connected with the purpose of denunciation -- an expressive function of punishment. Since prison symbolizes what it means to punish in our culture (Doob and Marinos 1995; Marinos 1995; Garland 1990), it is likely that some people will not view other ‘intermediate’ sanctions as effective and appropriate to accomplish this expressive purpose. Therefore one would expect that those who see denunciation as a highly important purpose of sentencing would be more likely to oppose interchangeability and support imprisonment.

Respondents were asked to rate the importance of a number of different purposes of sentencing, including denunciation (described as “expressing the community’s disapproval of the crime”), deterrence (“deterring the offender and other persons from committing offences”), incapacitation (“separating offenders from society”), rehabilitation (“assisting in the rehabilitation of offenders”), and compensation (“compensating victims or the community”). They were asked to rate the importance of these sentencing purposes on a scale where 1= not at all important, and 10= it is a very important purpose. Tables 2.13 and 2.14 present the findings on respondents ratings of denunciation and their willingness to support interchangeability of fines and community service orders for adult offenders.¹⁰

¹⁰ There were no similar findings on denunciation involving young offenders. Since denunciation is the focus of the analyses, the findings on other purposes of sentencing (including deterrence, incapacitation, rehabilitation, and compensation) are presented in an appendix to this Chapter.
Table 2.15

Respondents’ Views on the Importance of Denunciation as a Sentencing Purpose and Their Willingness to Support Interchangeability of a Fine for Adult Offenders

<table>
<thead>
<tr>
<th>Importance of Denunciation</th>
<th>Support for Substitution of Fine for Prison</th>
<th>Opposition to Substitution of Fine for Prison</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>“Low”</td>
<td>58.1% (86)</td>
<td>41.9% (62)</td>
<td>100% (148)</td>
</tr>
<tr>
<td>“Medium”</td>
<td>51.8% (85)</td>
<td>48.2% (79)</td>
<td>100% (164)</td>
</tr>
<tr>
<td>“High”</td>
<td>43.3% (71)</td>
<td>56.7% (93)</td>
<td>100% (164)</td>
</tr>
</tbody>
</table>

χ² = 6.93, df=2, p<.05

Table 2.16

Respondents’ Views on the Importance of Denunciation as a Sentencing Purpose and Their Willingness to Support Interchangeability of a Community Service Order for Adult Offenders

<table>
<thead>
<tr>
<th>Importance of Denunciation</th>
<th>Support for Substitution of CSO for Prison</th>
<th>Opposition to Substitution of CSO for Prison</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>“Low”</td>
<td>71.3% (107)</td>
<td>28.7% (43)</td>
<td>100% (150)</td>
</tr>
<tr>
<td>“Medium”</td>
<td>82.6% (138)</td>
<td>17.4% (29)</td>
<td>100% (167)</td>
</tr>
<tr>
<td>“High”</td>
<td>64.0% (105)</td>
<td>36.0% (59)</td>
<td>100% (164)</td>
</tr>
</tbody>
</table>

χ² = 14.69, df=2, p<0.001
The tables presented above (2.15 and 2.16) demonstrate that those who see denunciation as a highly important purpose of sentencing — for adult offenders only — are significantly more likely to oppose interchangeability, compared to those who do not see this function as important. There was no comparable effect for young offenders. While the overall pattern is demonstrated from “low” to “high” importance of denunciation, there is no apparent explanation for the peculiar pattern in the “medium” category of the importance of denunciation.

Thus it is evident from Tables 2.9 to 2.16 that support for interchangeability is not related to perceptions about the severity of sentencing practices in Canada for adult or young offenders. Rather, supporters of interchangeability — compared to those opposed — have broader related perceptions about punishment. Support and opposition to substitution of ‘intermediate’ sanctions is related to respondents’ perceptions about accountability of sanctions in the community, and beliefs in how the government should invest money in criminal justice matters. The findings reveal that respondents’ perceptions about sanctions in the community, and their belief in the prison — that it denounces criminal behaviour, and that the government should build more — are qualitative dimensions related to opposition of interchangeability.

---

11 A less surprising finding is the relationship between support for interchangeability and views about the importance of incapacitation. For both adult and young offenders, those respondents who view incapacitation as a highly important purpose of sentencing were significantly more likely to oppose interchangeability of a community service order, compared to those who see this purpose as having low importance. This relationship did not emerge for interchangeability of fines. These findings are presented in the appendix to this Chapter.
Conclusions

David Garland’s work (1990), as well as the work of others (Beattie 1986; Duff and Garland 1994), highlight a multidimensional approach to understanding the limits of people’s acceptance of different sanctions. Garland (1990) argues that we must understand punishment in a larger social context, rather than a focus on its criminal justice functions (e.g. the traditional goals of sentencing) (1990: 281-83). It is evident from this Chapter, however, that part of the social context is the act for which the offender is being punished, the nature of the ‘intermediate’ sanction that is to be imposed as an alternative to imprisonment, and who the offender is in terms of being an “adult” or “youth”.

The finding that sanctions are not equally substitutable for different offences (irrespective of the offender -- “adult” or “youth”) stands in contrast to Morris and Tonry’s (1990) conceptualization of interchangeability, and sentencing legislation for adults and youth. These authors do not suggest that there are limits to the appropriateness of interchangeability in terms of the nature of the offence, except for some offences that might be punishable by more than two years imprisonment. For example, they exclude some offences from interchangeability of fines and imprisonment: “[s]o perhaps, treason, murder, incest, rape, and bank robbery apart, it is not too difficult to invent cases of conviction of all other crimes appropriately punished by a fine, and certainly by a fine combined with another intermediate punishment” (Morris and Tonry 1990: 120). While they point to these inappropriate offences, their concerns are focused around issues of severity, rather than the qualities associated with the act being punished. Consistent with past research on interchangeability (Doob and Marinos 1995; Marinos 1997), this
Chapter revealed that members of the public were less supportive of the use of a fine or community service order for offences involving violence, compared to property offences.

The findings also illustrate some qualities about different sanctions that pose limits to public acceptance of interchangeability. Members of the public view the two ‘intermediate’ sanctions tested here as qualitatively different from each other, making one sanction more attractive for interchangeability than the other. In contrast, Morris and Tonry (1990) oversimplify the role of the “nature” of sanctions by conceptualizing community service orders in much the same way as fines. They state that

The prison is a punishment exacted against freedom of movement and association; the fine is a punishment exacted against money and what money can buy; the community service order is a punishment exacted against time and energy. Some see the community service order as a fine on time (1990: 150).

In this way, Morris and Tonry see these ‘intermediate’ sanctions as highly flexible and equally interchangeable with imprisonment -- along a single continuum -- so long as they are roughly equivalent in severity.

Sentencing legislation for adults and youth also does not provide guidance about whether one sanction might be more appropriate in one context more than another. However, it is likely that respondents in the present study associate more productive qualities to community service orders in terms of assisting both the offender and the community, compared to the use of fines. In fact some past research on public perceptions of community service orders reveals that members of the general public see community service orders in very favourable terms, and highly beneficial to both the offender and the community (Doob and Macfarlane 1984: 58-59).

Supporters of interchangeability, on the other hand, did not suggest the amount of
a fine or hours of community service in any consistent way, and thus the quantum of a punishment played a small role in their decision. One way that the quantitative aspects of interchangeability are addressed in practice is through the development of sentencing grids, found in a few American states. As mentioned in Chapter 1, sentencing grids outline the suggested penalty with the number of hours of community service or amount of a fine as substitutes for imprisonment when marked as appropriate, by focusing on the level of seriousness of the offence and the offender’s record.\textsuperscript{12} According to the results in this Chapter, however, it would be difficult to create a generic table where equivalence is accomplished without a consideration of the purposes of different punishments for different offences and offenders. Some who are willing to substitute might disagree with the severity of the ‘intermediate’ sanction, while those opposed altogether clearly view some sanctions as appropriate for some offences and not others -- a multidimensional perspective.

The nature of the offending population -- as an adult or youth -- was also considered by respondents when asked about their willingness to substitute a fine. What emerged was a difference in public support between fines and community service orders according to the nature of the offence. Fines were viewed as less appropriate for youth compared to adults, and community service orders were seen as more appropriate for youth compared to adults. It is likely that members of the public perceive youth as more capable of being accountable, and learning something positive from a community service order compared to paying a fine. A community service order is likely understood as involving hard work from which young offenders will learn about responsibility, while

\textsuperscript{12} For example, Morris and Tonry (1990) provide evidence of sentencing grids used in Minnesota and Washington state. In both grids, there are various levels of offences by criminal history of the offender
also symbolically "paying back" the community through service. A fine, on the other hand, might "hurt" an adult more than a youth because of their greater need for employment.

Exploring the two groups further -- supporters and opposers to interchangeability -- revealed that their preferences relate to broader beliefs about sentencing. Support for substituting a community service order for adult and young offenders was related to their views about investing in alternatives and crime prevention, perceptions about high completion rates of community service orders, and a relatively low importance placed upon denunciation as a sentencing principle. Those who support prisons instead of interchangeability, connect prison with what it means to punish. They prefer that more prisons are built and see prison as expressing denunciation of criminal behaviour. They also have negative views about a community sanction such as a community service order being completed. On the other hand, perceptions about the severity of sentences play a very small part in understanding support or opposition to interchangeability -- consistent with the other findings in this Chapter.

Overall, then, exploring public sensibilities about interchangeability reveals that punishment needs to be understood in more nuanced ways within the context of how sanctions are seen as accomplishing different purposes for different offences and offenders. These findings set the context for the following Chapter -- public sensibilities about the use of conditional sentences. Chapter 3 explores substitution of a conditional sentence of imprisonment for a term of imprisonment and the public's reactions.

which are eligible for interchangeability. See Chapter 3, "Interchangeability of Punishments in Practice."
Chapter 2: Appendix A

Respondents’ Views on the Importance of Each Sentencing Purpose
(Deterrence, Incapacitation, Rehabilitation, and Compensation)
and Their Willingness to Support Interchangeability
of a Fine or CSO for Adult Offenders

**Deterrence**

<table>
<thead>
<tr>
<th>D8BNEW IMPORT FOR SENT TO DETER by D18 Fine instead of prison</th>
</tr>
</thead>
<tbody>
<tr>
<td>Count</td>
</tr>
<tr>
<td>---</td>
</tr>
<tr>
<td>Row Pct</td>
</tr>
<tr>
<td>D8BNEW</td>
</tr>
<tr>
<td>low</td>
</tr>
<tr>
<td>high</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Column</td>
</tr>
<tr>
<td>Total</td>
</tr>
</tbody>
</table>

Chi-Square Value  DF  Significance
---  ---  ----
Pearson  .04382  1  .83418
Continuity Correction  .01366  1  .90695

Minimum Expected Frequency - 109.868
Number of Missing Observations: 31

**NOTE:**

- D8BNEW IMPORT FOR SENT TO DETER by D19 CSO instead of prison

- Chi-Square Value  DF  Significance
  ---  ---  ----
Pearson  .32090  1  .57107
Continuity Correction  .21391  1  .64372

Minimum Expected Frequency - 61.276
Number of Missing Observations: 25
### Incapacitation

**D8CNEW IMPORT FOR SENT TO SEPARATE OFFENDERS** by **D18** Fine instead of prison

<table>
<thead>
<tr>
<th>Count</th>
<th>Row Pct</th>
<th>Yes</th>
<th>No</th>
<th>Row Pct</th>
<th>Yes</th>
<th>No</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.00</td>
<td></td>
<td>129</td>
<td>115</td>
<td>2.00</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>low</td>
<td></td>
<td>52.9</td>
<td>47.1</td>
<td></td>
<td></td>
<td></td>
<td>51.6</td>
</tr>
<tr>
<td></td>
<td></td>
<td>112</td>
<td>117</td>
<td></td>
<td></td>
<td></td>
<td>229</td>
</tr>
<tr>
<td>high</td>
<td></td>
<td>48.9</td>
<td>51.1</td>
<td></td>
<td></td>
<td></td>
<td>49.4</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Column</th>
<th></th>
<th></th>
<th></th>
<th>Total</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>241</td>
<td>232</td>
<td></td>
<td>473</td>
<td>51.0</td>
<td>49.0</td>
<td>100.0</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Chi-Square</th>
<th>Value</th>
<th>DF</th>
<th>Significance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pearson</td>
<td>0.74147</td>
<td>1</td>
<td>0.38919</td>
</tr>
<tr>
<td>Continuity Correction</td>
<td>0.59146</td>
<td>1</td>
<td>0.44186</td>
</tr>
</tbody>
</table>

Minimum Expected Frequency - 112.321
Number of Missing Observations: 27

---

**D8CNEW IMPORT FOR SENT TO SEPARATE OFFENDERS** by **D19** Cso instead of prison

<table>
<thead>
<tr>
<th>Count</th>
<th>Row Pct</th>
<th>Yes</th>
<th>No</th>
<th>Row Pct</th>
<th>Yes</th>
<th>No</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.00</td>
<td></td>
<td>195</td>
<td>53</td>
<td>2.00</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>low</td>
<td></td>
<td>78.6</td>
<td>21.4</td>
<td></td>
<td></td>
<td></td>
<td>51.7</td>
</tr>
<tr>
<td></td>
<td></td>
<td>152</td>
<td>80</td>
<td></td>
<td></td>
<td></td>
<td>232</td>
</tr>
<tr>
<td>high</td>
<td></td>
<td>65.5</td>
<td>34.5</td>
<td></td>
<td></td>
<td></td>
<td>48.3</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Column</th>
<th></th>
<th></th>
<th></th>
<th>Total</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>147</td>
<td>133</td>
<td></td>
<td>480</td>
<td>72.3</td>
<td>27.7</td>
<td>100.0</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Chi-Square</th>
<th>Value</th>
<th>DF</th>
<th>Significance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pearson</td>
<td>10.28783</td>
<td>1</td>
<td>0.00134</td>
</tr>
<tr>
<td>Continuity Correction</td>
<td>9.64366</td>
<td>1</td>
<td>0.00190</td>
</tr>
</tbody>
</table>

Minimum Expected Frequency - 64.283
Number of Missing Observations: 20
## Rehabilitation

**D8DNEW IMPORT FOR SENT TO REHAB by D18 Fine instead of prison**

<table>
<thead>
<tr>
<th>D8DNEW</th>
<th>Yes</th>
<th>No</th>
</tr>
</thead>
<tbody>
<tr>
<td>low</td>
<td>136</td>
<td>131</td>
</tr>
<tr>
<td></td>
<td>50.9</td>
<td>49.1</td>
</tr>
<tr>
<td>high</td>
<td>110</td>
<td>101</td>
</tr>
<tr>
<td></td>
<td>52.1</td>
<td>47.9</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Row Pct</th>
<th>1.00</th>
<th>2.00</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total</td>
<td>246</td>
<td>232</td>
</tr>
<tr>
<td>Total</td>
<td>51.5</td>
<td>48.5</td>
</tr>
<tr>
<td>Total</td>
<td>100.0</td>
<td></td>
</tr>
</tbody>
</table>

Chi-Square Value: 0.06754, DF: 1, Significance: 0.79496

Continuity Correction: 0.02813, DF: 1, Significance: 0.86680

Minimum Expected Frequency: 102.410

Number of Missing Observations: 22

---

**D8DNEW IMPORT FOR SENT TO REHAB by D19 Cso instead of prison**

<table>
<thead>
<tr>
<th>D8DNEW</th>
<th>Yes</th>
<th>No</th>
</tr>
</thead>
<tbody>
<tr>
<td>low</td>
<td>185</td>
<td>84</td>
</tr>
<tr>
<td></td>
<td>68.8</td>
<td>31.2</td>
</tr>
<tr>
<td>high</td>
<td>168</td>
<td>47</td>
</tr>
<tr>
<td></td>
<td>78.1</td>
<td>21.9</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Row Pct</th>
<th>1.00</th>
<th>2.00</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total</td>
<td>353</td>
<td>131</td>
</tr>
<tr>
<td>Total</td>
<td>72.9</td>
<td>27.1</td>
</tr>
<tr>
<td>Total</td>
<td>100.0</td>
<td></td>
</tr>
</tbody>
</table>

Chi-Square Value: 5.31039, DF: 1, Significance: 0.02120

Continuity Correction: 4.84651, DF: 1, Significance: 0.02770

Minimum Expected Frequency: 58.192

Number of Missing Observations: 16
Compensation

**D8NEW IMPORT FOR SENT TO COMPENSATE VIC by D18 Fine instead of prison**

<table>
<thead>
<tr>
<th></th>
<th>Yes</th>
<th>No</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>D8NEW</strong></td>
<td>1.00</td>
<td>139</td>
<td>116</td>
</tr>
<tr>
<td></td>
<td>2.00</td>
<td>107</td>
<td>118</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>246</td>
<td>234</td>
<td>480</td>
</tr>
</tbody>
</table>

Chi-Square Table:

<table>
<thead>
<tr>
<th>Chi-Square</th>
<th>Value</th>
<th>DF</th>
<th>Significance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pearson</td>
<td>2.31373</td>
<td>1</td>
<td>.12824</td>
</tr>
<tr>
<td>Continuity Correction</td>
<td>2.04376</td>
<td>1</td>
<td>.15283</td>
</tr>
</tbody>
</table>

Minimum Expected Frequency: 109.688
Number of Missing Observations: 20

---

**D8NEW IMPORT FOR SENT TO COMPENSATE VIC by D19 Cso instead of prison**

<table>
<thead>
<tr>
<th></th>
<th>Yes</th>
<th>No</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>D8NEW</strong></td>
<td>1.00</td>
<td>198</td>
<td>60</td>
</tr>
<tr>
<td></td>
<td>2.00</td>
<td>154</td>
<td>73</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>352</td>
<td>133</td>
<td>485</td>
</tr>
</tbody>
</table>

Chi-Square Table:

<table>
<thead>
<tr>
<th>Chi-Square</th>
<th>Value</th>
<th>DF</th>
<th>Significance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pearson</td>
<td>4.80888</td>
<td>1</td>
<td>.02831</td>
</tr>
<tr>
<td>Continuity Correction</td>
<td>4.37197</td>
<td>1</td>
<td>.03653</td>
</tr>
</tbody>
</table>

Minimum Expected Frequency: 62.249
Number of Missing Observations: 15
Chapter 2: Appendix B

Respondents' Views on the Importance of Each Sentencing Purpose
(Deterrence, Incapacitation, Rehabilitation, and Compensation)
and Their Willingness to Support Interchangeability
of a Fine or CSO for Young Offenders

**Deterrence**

<table>
<thead>
<tr>
<th>D8BNEW IMPORT FOR SENT TO DETER by D18 Fine instead of prison</th>
</tr>
</thead>
<tbody>
<tr>
<td>Count</td>
</tr>
<tr>
<td>-------</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>D8BNEW</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>low</td>
</tr>
<tr>
<td>high</td>
</tr>
<tr>
<td>Column</td>
</tr>
<tr>
<td>Total</td>
</tr>
</tbody>
</table>

Chi-Square Value | DF | Significance
-----------------|----|---------------
Pearson          | 2.59928 | 1 | .10691
Continuity Correction | 2.31202 | 1 | .12838
Minimum Expected Frequency | 95.791 |
Number of Missing Observations: 17

<table>
<thead>
<tr>
<th>D8BNEW IMPORT FOR SENT TO DETER by D19 CSO instead of prison</th>
</tr>
</thead>
<tbody>
<tr>
<td>Count</td>
</tr>
<tr>
<td>-------</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>D8BNEW</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>low</td>
</tr>
<tr>
<td>high</td>
</tr>
<tr>
<td>Column</td>
</tr>
<tr>
<td>Total</td>
</tr>
</tbody>
</table>

Chi-Square Value | DF | Significance
-----------------|----|---------------
Pearson          | .05188 | 1 | .81982
Continuity Correction | .01339 | 1 | .90788
Minimum Expected Frequency | 45.984 |
Number of Missing Observations: 14
Incapacitation

D8CNEW IMPORT FOR SENT TO SEPARATE OFFENDERS by D18 Fine instead of prison

<table>
<thead>
<tr>
<th></th>
<th>Count</th>
<th>D18</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Yes</td>
<td></td>
<td>No</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Row Pct</td>
<td></td>
<td>1.00</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>low</td>
<td>1.00</td>
<td>120</td>
<td></td>
<td>145</td>
<td>265</td>
</tr>
<tr>
<td></td>
<td></td>
<td>45.3</td>
<td></td>
<td>54.7</td>
<td>54.8</td>
</tr>
<tr>
<td>high</td>
<td>2.00</td>
<td>91</td>
<td></td>
<td>128</td>
<td>219</td>
</tr>
<tr>
<td></td>
<td></td>
<td>41.6</td>
<td></td>
<td>58.4</td>
<td>45.2</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Column</td>
<td>211</td>
<td>273</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Total</td>
<td>436</td>
<td>564</td>
</tr>
</tbody>
</table>

Chi-Square

<table>
<thead>
<tr>
<th>Value</th>
<th>DF</th>
<th>Significance</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pearson</td>
<td>.67862</td>
<td>1</td>
</tr>
<tr>
<td>Continuity Correction</td>
<td>.53539</td>
<td>1</td>
</tr>
<tr>
<td>Minimum Expected Frequency</td>
<td>95.473</td>
<td></td>
</tr>
<tr>
<td>Number of Missing Observations: 22</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

D8CNEW IMPORT FOR SENT TO SEPARATE OFFENDERS by D19 Cso instead of prison

<table>
<thead>
<tr>
<th></th>
<th>Count</th>
<th>D19</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Yes</td>
<td></td>
<td>No</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Row Pct</td>
<td></td>
<td>1.00</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>low</td>
<td>1.00</td>
<td>226</td>
<td></td>
<td>41</td>
<td>267</td>
</tr>
<tr>
<td></td>
<td></td>
<td>84.6</td>
<td></td>
<td>15.4</td>
<td>54.9</td>
</tr>
<tr>
<td>high</td>
<td>2.00</td>
<td>159</td>
<td></td>
<td>60</td>
<td>219</td>
</tr>
<tr>
<td></td>
<td></td>
<td>72.6</td>
<td></td>
<td>27.4</td>
<td>45.1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Column</td>
<td>385</td>
<td>101</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Total</td>
<td>79.2</td>
<td>20.8</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>100.0</td>
</tr>
</tbody>
</table>

Chi-Square

<table>
<thead>
<tr>
<th>Value</th>
<th>DF</th>
<th>Significance</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pearson</td>
<td>10.59662</td>
<td>1</td>
</tr>
<tr>
<td>Continuity Correction</td>
<td>9.87782</td>
<td>1</td>
</tr>
<tr>
<td>Minimum Expected Frequency</td>
<td>45.512</td>
<td></td>
</tr>
<tr>
<td>Number of Missing Observations: 20</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Rehabilitation

**D8DNEW IMPORT FOR SENT TO REHAB by D18 Fine instead of prison**

<table>
<thead>
<tr>
<th>D8DNEW</th>
<th>Count</th>
<th></th>
<th></th>
<th>Row</th>
<th></th>
<th></th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Yes</td>
<td>No</td>
<td></td>
<td></td>
<td>1.00</td>
<td>2.00</td>
<td></td>
</tr>
<tr>
<td>low</td>
<td>104</td>
<td>145</td>
<td></td>
<td>249</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>41.8</td>
<td>58.2</td>
<td></td>
<td>50.5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>high</td>
<td>111</td>
<td>133</td>
<td></td>
<td>244</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>45.5</td>
<td>54.5</td>
<td></td>
<td>49.5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Column</td>
<td>215</td>
<td>278</td>
<td></td>
<td>493</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>43.6</td>
<td>56.4</td>
<td></td>
<td>100.0</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Chi-Square

<table>
<thead>
<tr>
<th>Value</th>
<th>DF</th>
<th>Significance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pearson</td>
<td>.69525</td>
<td>1</td>
</tr>
<tr>
<td>Continuity Correction</td>
<td>.55204</td>
<td>1</td>
</tr>
</tbody>
</table>

Minimum Expected Frequency - 106.410
Number of Missing Observations: 13

**D8DNEW IMPORT FOR SENT TO REHAB by D19 Cso instead of prison**

<table>
<thead>
<tr>
<th>D8DNEW</th>
<th>Count</th>
<th></th>
<th></th>
<th>Row</th>
<th></th>
<th></th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Yes</td>
<td>No</td>
<td></td>
<td></td>
<td>1.00</td>
<td>2.00</td>
<td></td>
</tr>
<tr>
<td>low</td>
<td>197</td>
<td>54</td>
<td></td>
<td>251</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>78.5</td>
<td>21.5</td>
<td></td>
<td>50.6</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>high</td>
<td>199</td>
<td>46</td>
<td></td>
<td>245</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>81.2</td>
<td>18.8</td>
<td></td>
<td>49.4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Column</td>
<td>396</td>
<td>100</td>
<td></td>
<td>496</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>79.8</td>
<td>20.2</td>
<td></td>
<td>100.0</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Chi-Square

<table>
<thead>
<tr>
<th>Value</th>
<th>DF</th>
<th>Significance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pearson</td>
<td>.57760</td>
<td>1</td>
</tr>
<tr>
<td>Continuity Correction</td>
<td>.42001</td>
<td>1</td>
</tr>
</tbody>
</table>

Minimum Expected Frequency - 49.395
Number of Missing Observations: 10
## Compensation

**DBENEW IMPORT FOR SENT TO COMPENSATE VIC by D18 Fine instead of prison**

<table>
<thead>
<tr>
<th></th>
<th>Yes</th>
<th>No</th>
<th>Row</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>DBENEW</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>low</td>
<td>117</td>
<td>150</td>
<td>267</td>
</tr>
<tr>
<td></td>
<td>43.8</td>
<td>56.2</td>
<td>54.5</td>
</tr>
<tr>
<td>high</td>
<td>96</td>
<td>127</td>
<td>223</td>
</tr>
<tr>
<td></td>
<td>43.0</td>
<td>57.0</td>
<td>45.5</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>213</td>
<td>277</td>
<td>490</td>
</tr>
<tr>
<td><strong>Chi-Square Value</strong></td>
<td>0.2939</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Continuity Correction</strong></td>
<td>0.0639</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Minimum Expected Frequency</strong></td>
<td>96.937</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Number of Missing Observations</strong></td>
<td>16</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**DBENEW IMPORT FOR SENT TO COMPENSATE VIC by D19 Cso instead of prison**

<table>
<thead>
<tr>
<th></th>
<th>Yes</th>
<th>No</th>
<th>Row</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>DBENEW</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>low</td>
<td>220</td>
<td>49</td>
<td>269</td>
</tr>
<tr>
<td></td>
<td>81.8</td>
<td>18.2</td>
<td>54.7</td>
</tr>
<tr>
<td>high</td>
<td>170</td>
<td>53</td>
<td>223</td>
</tr>
<tr>
<td></td>
<td>76.2</td>
<td>23.8</td>
<td>45.3</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>390</td>
<td>102</td>
<td>492</td>
</tr>
<tr>
<td><strong>Chi-Square Value</strong></td>
<td>2.28629</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Continuity Correction</strong></td>
<td>1.96097</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Minimum Expected Frequency</strong></td>
<td>46.232</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Number of Missing Observations</strong></td>
<td>14</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Chapter 3

Public Sensibilities about Conditional Sentences of Imprisonment

Chapter 2 examined public sensibilities about the interchangeability of fines and community service orders for prison sentences. The findings revealed that substituting these sanctions for sentences of imprisonment are seen by members of the public as limited in some circumstances involving violence, and for some offenders (such as community service orders for adult offenders). Opposers of interchangeability, compared to supporters, saw denunciation as a significantly more important purpose of sentencing, and advocated investing in prisons. Chapter 3 addresses public sensibilities about the conditional sentence of imprisonment.¹ In this Chapter, questions focus on the range of their support and opposition towards this sanction for different offences, and the possible reasons for their views. Consistent with the findings in the previous Chapter, it is likely that public perceptions of the appropriateness of conditional sentences is more complex than assumed.

Within the criminal law community, the conditional sentence is a controversial sanction. Courts of Appeal have given it a fair amount of attention and the Supreme Court of Canada has heard a package of appeals on various issues concerning whether it is being used appropriately.² On the other hand, when the conditional sentence was introduced, as part of the package of amendments to the Criminal Code dealing with sentencing matters, it received very little public attention during its passage through

---

¹ This Chapter is based on a paper by V. Marinos and A. N. Doob. 1999. “Understanding Public Attitudes Toward Conditional Sentences of Imprisonment.” 21 Criminal Reports (5th) 31-41.
Parliament. This was largely due to the fact that Parliament focused, in its wisdom, on the implications of including two words -- sexual orientation -- in a list of "hate motivated" crimes. Thus, for the public, the bill known to the legal community as Bill C-41 and seen as addressing a long list of sentencing issues, was often described in the mass media as the "gay rights" bill.

The result of this is that when conditional sentences started being handed down in courts -- and Courts of Appeal began addressing the complexities of this new sanction -- the public was faced with the task of interpreting sentences which contained apparently contradictory sounding terms. Although Courts of Appeal have different views of what the proper process is by which a conditional sentence should be imposed, conditional sentences are typically described as "sentences of imprisonment" where the judge "then" allows the offender to serve the sentence in the community. If the public might find the conditional sentence difficult to understand, they are not alone. Though the conditional sentence might have had an easy birth, Parliament appears not to have had enormous confidence in the details of its new sanction. The sanction has already received two corrective operations.

We were interested in how this new sanction was being perceived by the public in its early days. However, because at the time that the present study was being carried out (Summer 1997) the conditional sentence was relatively unknown, we decided it was

---

S.C.C. 5.


4 The Alberta Court of Appeal in Brady appears to take the position that the judge should not go through the "two step" process described here, but instead should consider the conditional sentence to be, in effect, just another intermediate sanction. In Proud, the Supreme Court of Canada did not agree with the Alberta Court of Appeal.

5 Section 742.1(b) of the Code added the requirement that conditional sentences of imprisonment must be consistent with the purposes and principles of sentencing. See section 742.4(6) of the Code modified some
inappropriate simply to ask questions about it. Rather we described the conditional sentence in a summary form and examined the patterns of support and opposition to the sanction. One advantage of looking at the public’s response to conditional sentences early in the process is that it should be possible to see the reaction of members of the public relatively uncontaminated by inadequate information and inflammatory statements by its critics conveyed in mass media. The result was that during the summer of 1997, 500 adult residents of Ontario, representative of adults in the province, were contacted by telephone and were asked a series of questions having to do with crime and the

of the conditions related to breaches of conditional sentence orders.

6A conditional sentence of imprisonment was described throughout the survey as follows: “A conditional sentence is one where the offender is sentenced to prison for a period of up to two years, but instead of prison he is able to serve it in the community, if it is believed he is not a danger to the community. Various specific conditions can be placed on the offender such as not being able to leave his home except to go to work, abstaining from alcohol, attending a drug, alcohol or psychiatric treatment program, doing a specified number of hours of community work, etc. If the offender is found to have violated any of these conditions, he then has to serve the remainder of the sentence in prison.” Any definition of a complex sanction such as this is arbitrary. In a survey a compromise has to be arrived at between being “legally correct” on the one hand, and understandable on the other. For example, the consequences of a finding that the offender breached are not as simple as described here. By indicating that a breach would result in imprisonment we may have described the sanction in more favourable terms for those who favour imprisonment. It is very unlikely, however, that minor details such as this would affect the pattern of results reported here.

7In one case, where a police officer found guilty of criminal negligence causing death was given a conditional sentence it was described in one news article as a sentence of “two years less a day to be served in the community. He must perform 180 hours of unspecified community service and can’t possess a weapon until his sentence has been served” (Freed, D. A. “Officer Could be Charged Under Police Act.” The Toronto Star, Saturday July 5, 1997, A9). In a column appearing three days later in the same newspaper, the writer says that the offender was “sentenced to community service” (Urquhart, I. “Legal Smoke Screen Fogs Ipperwash Affair.” The Toronto Star, July 8, 1997, A15). In a different case, a story indicates that an offender received “a one-year conditional sentence” but the headline read “Man Who Shot Dead Best Friend Escapes Further Time in Jail” (The Toronto Star, Saturday, April 19, 1997).

8 Any survey result is obviously an estimate of what the “true” population value (in this case adult Ontario residents) would be. Assuming that we do, in fact, have a “representative” sample of Ontario residents, we can estimate how likely it is that the true population value is within a certain range of possible scores (the “confidence limits”). In the case of the findings presented here with a sample of 500 people, we would expect that the true population percentage would be within about 4.4% of the obtained percentage in 95% of the estimates that are presented. Note, however, that for many of our figures, our sample is considerably smaller. Thus for the first row of data in Table 3.1, the 95% confidence limits (the limits within which one would reasonably expect the true proportion to lie, 95% of the time) are 7.5%. Largely, however, confidence limits are irrelevant for most of the comparisons we make since we are, in fact, talking about comparisons between groups (e.g., groups assessing different cases). In these cases, we are more interested in whether the difference between two groups is reliable (“statistically significant”) not in estimating the exact percentage of people favouring a particular outcome.
Support for the conditional sentence

Because the survey that we carried out was not designed to "map" the full breadth of support or opposition to the use of conditional sentences, we cannot give a definitive answer to the question, "When are conditional sentences seen as being appropriate?" What we can say, however, is that there are clearly circumstances where many people see conditional sentences as being entirely appropriate sanctions for certain offences involving violence.

One question in the survey asked respondents to indicate their views of whether a person should be given an ordinary prison sentence or a conditional sentence for one of the following cases:

- A 23 year old man who had been found guilty in four separate incidents where he and his friends broke into houses and stole various items.
- A 23 year old man who had been found guilty of the sexual assault of a 22 year old woman he had known for a couple of years.
- A 23 year old man who had been found guilty of assault causing bodily harm after he hit and broke the nose of a man he had a disagreement with at a local bar.

All respondents were told that "A judge decides that the appropriate sentence

---

The survey was carried out by Goldfarb Consultants using standard commercial telephone survey methodology. Residents were contacted using a random digit dialing technique such that the numbers chosen were representative of Ontario households. One adult living in each household that was contacted was asked to respond to the survey. Detailed findings of the findings related to conditional sentences are contained in V. Marinos and A. N. Doob A Preliminary Examination of Public Views of Conditional Sentences of Imprisonment (A report to the Department of Justice, Canada, 1997). Findings from the full survey (including a parallel survey dealing with youth crime and the youth justice system) are contained in A. Doob, J. Sprott, V. Marinos, and K. N. Varma. An Exploration of Ontario Residents' Views of Crime and the Criminal Justice System (Toronto: Centre of Criminology, 1998).
should be imprisonment for 18 months. The judge is satisfied, however, that if the offender were to serve his sentence in the community, he would not endanger the safety of the community. The judge has not yet decided whether the sentence should be served in a prison, or as a conditional sentence of imprisonment." The description of release rules for prison as well as the meaning of a conditional sentence were then read to respondents and they were asked to indicate their choice. The results are shown in Table 3.1.

Table 3.1

Preference for Prison or Conditional Sentence as a Function of Offence

<table>
<thead>
<tr>
<th>Offence:</th>
<th>Respondent’s preference for this case:</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Normal Prison</td>
<td>Conditional Sentence</td>
</tr>
<tr>
<td>Break &amp; Enter</td>
<td>55.6%</td>
<td>44.4%</td>
</tr>
<tr>
<td>(n=169)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sexual Assault</td>
<td>60.3%</td>
<td>39.7%</td>
</tr>
<tr>
<td>(n=141)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Assault Bodily</td>
<td>28.9%</td>
<td>71.1%</td>
</tr>
<tr>
<td>Harm</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\(\chi^2 = 38.72, \text{df}=2, p<.001\)

There are two points that should be made concerning the results described in Table 3.1. First of all, it is clear that a substantial number of people for all offences preferred a conditional sentence to a sentence served in a prison. Second, while in all cases we were starting from a "standard" eighteen month sentence where the offender was described as not being a threat to the community, the preference for a conditional sentence varied considerably across the three offences. In particular, it is interesting to note that it was in the case of the assault that the conditional sentence was most attractive as an alternative to prison. And, given the controversy as to whether conditional
sentences should be given in cases of sexual assault\(^\text{10}\) it is interesting to note that the willingness to substitute a conditional sentence for a normal prison sentence was not significantly less in the case of the sexual assault compared to the break and enter case.

Part of the appeal of a conditional sentence -- to an offender as well as to the public when contemplating a sentence of a particular length -- is that it may be seen as reducing the harshness of a sentence of imprisonment. The question that arises, then, is whether it is seen as being any different from a term of probation.

Respondents were told that a 23 year old was found guilty of breaking into a hardware store and taking goods worth about $800. They were told he had a record of similar types of offences. The sentence that the judge imposed was described as being one of the following:

- A prison sentence (release provisions were described)
- A conditional sentence of imprisonment (described as in the earlier case)
- A 12 month term of probation (the meaning of which was described to the respondents)

Survey respondents were asked to indicate\(^\text{11}\) how appropriate they found the sentence that was described to them. As can be seen in Table 3.2, there was a good deal of variation in the ratings of the appropriateness of the sentence across cases. In particular, it would appear that a prison sentence was seen as being inappropriate probably because it was seen to be too harsh. In this particular instance, then, a

---


\(^{11}\) The respondents were asked to answer on a ten point scale running from “completely inappropriate” to “completely appropriate.” The results can be presented in a number of different ways, but the overall pattern is the same as in Table 3.2. In Table 3.2, we divided the scale into two, calling scale points 1-5 “inappropriate” and 5-10 “appropriate.”
conditional sentence was clearly seen as being different from (and more appropriate than) a prison sentence. However, it is noteworthy that respondents did not see it as different from a term of probation.

Table 3.2

Appropriateness of a Sentence of Prison, Conditional Sentence, or Probation for a Break and Enter into a Commercial Establishment by a Man with a Criminal Record

<table>
<thead>
<tr>
<th>Sentence:</th>
<th>Inappropriate</th>
<th>Appropriate</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prison</td>
<td>58.9%</td>
<td>41.1%</td>
<td>100% (n=151)</td>
</tr>
<tr>
<td>Conditional sentence</td>
<td>35.8%</td>
<td>64.2%</td>
<td>100% (n=176)</td>
</tr>
<tr>
<td>Probation</td>
<td>44.6%</td>
<td>55.4%</td>
<td>100% (n=168)</td>
</tr>
</tbody>
</table>

χ² = 17.69, df=2, p<.001

Clearly the overall concept of a conditional sentence -- that convicted offenders have prison sentences hanging over their heads which can result in a shift from the community to a "real" prison if they do not follow certain conditions -- is not only acceptable to many people but it is also preferred in some moderately serious cases. However, the results described in Table 3.2 suggest that what may be important about a conditional sentence is simply that the offender does not immediately go to prison. The result is that a conditional sentence is not viewed as being different from probation.
Is opposition to conditional sentences associated with the view that sentences are too lenient?

Although it is argued by some that the conditional sentence of imprisonment is little more than a “soft option” instead of imprisonment,\(^{12}\) it is clearly meant to be -- and meant to be seen to be -- a “type” of prison sentence rather than just another ‘intermediate’ sanction. Opposition to it, then, may come primarily from those who view sentences generally as being too lenient. In this sample, as in most surveys carried out in Canada and the US over the past thirty years,\(^{13}\) most people thought that sentences were not severe enough (in this survey about 77% expressed this belief). If opposition to imposing conditional sentences stems from a view that these sentences are too lenient, then those who hold the belief that all adult sentences are too lenient should be much more likely to oppose the use of conditional sentences compared to those who believe that adult sentences are about right or too severe.

As shown in Table 3.3, those who think that sentences are too lenient are slightly less favourable toward conditional sentences. However, this effect is not statistically significant and, therefore, should best be considered to be random variation. In other words, those respondents who indicated that they believed that sentences, generally, were too lenient, were no more likely than those who indicated that they were content with current sentencing severity to favour a normal prison sentence over a conditional sentence.

---


Table 3.3

Preference for a Normal Prison Sentence or a Conditional Sentence as a Function of Respondents’ Views of the Severity of Adult Sentences

<table>
<thead>
<tr>
<th>Adult sentences are:</th>
<th>Normal Prison</th>
<th>Conditional Sentence</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>About right (^{14})</td>
<td>39.0%</td>
<td>61.0%</td>
<td>100% (n=100)</td>
</tr>
<tr>
<td>Not severe enough</td>
<td>48.9%</td>
<td>51.1%</td>
<td>100% (n=366)</td>
</tr>
</tbody>
</table>

\(\chi^2\) (corrected) = 2.71, df =1, \( p = .10\), not significant. Note: These results are pooled across the three offences.

Support for prisons

We asked two questions that addressed directly the support respondents had for the use of prisons. Both questions were prefaced by a statement that the province’s prisons are full. The first question asked respondents to choose between two solutions: build more prisons or sentence more people to alternatives to custody (some examples of these were included). The second question reminded respondents that the province’s prisons were full and asked them whether they would favour building more prisons or investing in programs to prevent crime.

Building more prisons was not seen, by most Ontario residents as an attractive policy option either when pitted against increased use of alternatives (More prisons: 34.5%; More use of alternatives: 65.5%) or crime prevention (More prisons: 14%; More crime prevention: 86%). In the context of this paper, however, what is important is that those who favoured prison -- rather than alternatives to prison (Table 3.4) or crime

---

\(^{14}\)The very small number of people who indicated that they thought that sentences were "too severe" are pooled with those who indicated that adult sentences were "about right."
prevention (Table 3.5) -- were much less likely to favour the 18 month conditional sentence over the 18 month conventional prison sentence.

Table 3.4

Preference for a Normal Prison Sentence or a Conditional Sentence as a Function of Respondents' Preference for More Prisons or More Use of Alternatives to Custody

<table>
<thead>
<tr>
<th>Respondent prefers:</th>
<th>Normal Prison</th>
<th>Conditional Sentence</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>More Prisons</td>
<td>61.5%</td>
<td>38.5%</td>
<td>100%</td>
</tr>
<tr>
<td>(n=161)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>More Alternatives</td>
<td>38.9%</td>
<td>61.1%</td>
<td>100%</td>
</tr>
<tr>
<td>(n=301)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

$\chi^2$ (corrected) = 21.56; df=1, $p<.001$. Note: These results are pooled across the three offences.

Table 3.5

Preference for a Normal Prison Sentence or a Conditional Sentence as a Function of Respondents' Preference for More Prisons or Investment in Crime Prevention

<table>
<thead>
<tr>
<th>Respondent prefers:</th>
<th>Normal Prison</th>
<th>Conditional Sentence</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>More Prisons</td>
<td>67.6%</td>
<td>32.4%</td>
<td>100%</td>
</tr>
<tr>
<td>(n=68)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Crime Prevention</td>
<td>42.9%</td>
<td>57.1%</td>
<td>100%</td>
</tr>
<tr>
<td>(n=410)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

$\chi^2$ (corrected) = 13.35; df=1, $p<.00$. Note: These results are pooled across the three offences.

Clearly, opposition to the use of conditional sentences is based, in part, on support for prisons -- presumably a belief in the ability of prisons to achieve something. Opposition to conditional sentences does not, however, appear to relate to beliefs about sentence severity per se. That is, it is not simply that imprisonment is seen as the most
severe sanction and is, therefore, supported. Hence we must look further for an explanation. We asked all respondents to indicate\textsuperscript{15} the importance of each of five purposes of sentencing: expressing the community’s disapproval of the crime, deterring the offender and others from committing offences, separating offenders from society, assisting in the rehabilitation of offenders, and compensating victims or the community. Only one of these purposes related to the choice of a conditional sentence over imprisonment: those respondents who rated “separating offenders from society” higher were less likely (than those who saw incapacitation as being less important) to favour a conditional sentence over a normal prison sentence. Of those who saw incapacitation as being very important, 47% favoured a conditional sentence. Of those who rated incapacitation as being somewhat less important, 58.1% favoured a conditional sentence.\textsuperscript{16}

In addition, Courts of Appeal\textsuperscript{17} and public attitude surveys\textsuperscript{18} suggest that where denunciation is seen as being important in a case, “intermediate” sanctions such as fines and community service orders are seen as being inferior to prison in accomplishing this goal of punishment. Thus it is likely that prison is viewed as symbolic of what it means to punish and to express condemnation of the behaviour.\textsuperscript{19}

\textsuperscript{15}Respondents used a 10 point scale going from 1= Not at all important to 10= very important to rate the importance of a number of traditional sentencing goals. Because respondents tended to rate all purposes as being at least reasonably important, (i.e., most were near to top of the 10 point scale) we divided the sample into two groups (high and low) at the median on each dimension. It should be remembered that “low” in this case is only relative.

\textsuperscript{16}Chi-square, corrected = 5.47, df=1, p <.02.

\textsuperscript{17}e.g. Wismayer (1997) 115 C.C.C. (3d) 18 (Ont. C.A.)


Conclusion

In general, it is clear that a conditional sentence of imprisonment is seen by many people as a viable alternative to a normal sentence of imprisonment. To the extent that its purpose is largely to provide another sentencing option between doing essentially nothing and sending the offender to prison, it fits the description of an "intermediate sanction" described, and advocated, by Morris and Tonry\(^\text{20}\) and others. The difficulty as revealed in Table 3.2, is that this "new" sanction may not be seen as any different from a probation sentence. If this is the case, its usefulness is somewhat limited except as a way of giving sentencing judges some choices when faced with firm guidance from their Courts of Appeal.\(^\text{21}\)

However, as suggested by other research on other sanctions\(^\text{22}\) it may be that the conditional sentence of imprisonment is not viewed by the public as being equally acceptable for all offences. Elsewhere\(^\text{23}\) we have noted that conditional sentences may be seen as being an appropriate way of accomplishing certain sentencing purposes for some offences, but not for others. Fines, community service orders, and conditional sentences of imprisonment all varied in their perceived appropriateness as a means to accomplish

---


\(^{21}\)See Manson, A. "The Reform of Sentencing in Canada." Paper prepared for conference on "Making Criminal Law Clear and Just" (November 1998, Queen's University; Kingston, Ontario) who argues (p. 29) that conditional sentences may serve the function of allowing trial judges to have their cake and eat it, too, by imposing "prison" sentences where their court of appeal has said that one is necessary but at the same time not sending a person to prison where a prison sentence makes no sense. See also Manson, A. "Finding a Place for Conditional Sentences." (1997) 3 \textit{Criminal Reports} (5th) 183.


\(^{23}\)Marinos, V. and A.N. Doob. "A Preliminary Examination of Public Views of Conditional Sentences of Imprisonment A Report to the Department of Justice, Canada." Centre of Criminology, University of
certain purposes at sentencing depending on the offence.

It has been argued\(^\text{24}\) that the public will see a conditional sentence as an instance of the offender being "let off" easily. Obviously this is only the case if the sentence is meant to be interpreted as being equivalent, in severity, to a normal prison sentence. If, on the other hand, it is seen simply as one more intermediate sanction, then a different set of concerns arise. What is clear from our data, however, is that a conditional sentence is not seen, by those in opposition to it, as accomplishing the same purposes as imprisonment. Hence attempts to create more onerous conditions for the conditional sentence will not solve the problem for those who appear to "like" prisons unless, of course, conditions are imposed which are seen to be not only as onerous, but also as effective in separating the offender from society.

Courts of Appeal, like members of the public and legal commentators, are divided on the usefulness of conditional sentences of imprisonment. Those who seem to be less in favour of the heavy use of imprisonment may be more inclined to try to find a place for the conditional sentence in the sentencing structure than are those who, for various reasons, believe in the efficacy of imprisonment.

---

Chapter 4

What’s Intermediate about ‘Intermediate’ Sanctions? The Combination of ‘Intermediate’ Sanctions and Denunciatory Sentences in the Courts

Chapters 2 and 3 have examined public sensibilities about interchangeability within the context of perceived connections made among the purposes of sentencing, sanctions, and the nature of the offence and the offender. The purpose of denunciation has been found to be a significant factor in people’s perceptions about the appropriateness of substituting sanctions for prison sentences for some violent offences. The present Chapter focuses on the combination of ‘intermediate’ sanctions and lengths of prison sentences handed out by the courts. Using youth court dispositions for 1994-95, this analysis presents findings on the combinations of sanctions imposed, and in particular, the reasons for the combination of ‘intermediate’ sanctions along with short periods of custody for particular offences.\(^1\) This study challenges simplistic conceptualizations of punishment — that sanctions differ solely on a dimension of severity, and that sanctions are highly fluid and flexible entities which can easily be calibrated into equivalencies at sentencing.

The concept of ‘intermediate’ sanctions\(^2\) has become increasingly popular in the literature on punishment. Intermediacy in punishment has been defined as a range of sanctions between the two extreme poles of probation and imprisonment. Recent strategies of penal reform in the United States and Canada have called for an increased use of ‘intermediate’ sanctions (Morris and Tonry 1990; Byrne et al. 1992; Smykla

---


\(^2\)The research and preparation of this paper were supported by a Social Sciences and Humanities Research
Selke 1995; Tonry and Hamilton 1995). For example, Morris and Tonry (1990) argue that in the United States, probation and prison have become overused sanctions in the 1980s and 1990s, and that the fiscal and prison crowding crises require attention. Thus what is needed, they argue, is a range of meaningful sanctions between probation and prison (Morris and Tonry 1990: 3).

Intermediacy in punishment has become an important penal strategy in Canada. First, sentencing legislation requires that judges use imprisonment only when necessary for both adults and young offenders. For adults, amendments to the Criminal Code which came into effect in September 1996 specify that sanctions other than imprisonment should be considered for all offenders (sections 718.2 (c) & (d)). In sentencing young offenders, the fundamental principles of the Young Offenders Act provide that judges should impose the least restrictive means possible in the circumstances. Thus judges are to consider numerous ways of dealing with the youth other than imprisonment (see section 24 (1.1) (b) and (c)). Second, we have also seen the development of new sanctions such as conditional sentences of imprisonment for adult offenders (section 724 of the Criminal Code) which are consistent with the rationale of ‘intermediacy’. This analysis explores the imposition of ‘intermediate’ sanctions for young offenders in Canadian youth courts for 1994-5, and suggests the purpose of combining an ‘intermediate’ sanction and custody.

The Imprisonment-Denunciation Nexus

While decreasing the use of imprisonment and increasing the use of ‘intermediate’ sanctions are important goals in sentencing, it is critical to inquire into the

Council of Canada grant to A.N. Doob.
sensibilities and purposes of different sanctions. Analyses of penal changes over time suggest that the values and meanings underlying sanctions have led to the development and demise of particular penal practices (Beattie 1986; Braithwaite 1989). Doob and Marinos (1995) argue that different sanctions have different meanings across time and place and serve a variety of purposes. Thus they suggest that certain ‘intermediate’ sanctions, though punitive, may not be seen as capable of serving particular purposes for some offences.

Proposals to replace imprisonment with ‘intermediate’ sanctions have not considered adequately the broader purposes of different sanctions in different contexts. It is assumed that fines and community service orders, for instance, are highly flexible and can be made equivalent to imprisonment for a variety of offences by increasing the size of these sanctions (Morris and Tonry 1990; Smykla and Selke 1995). In fact there is some evidence to suggest that the meanings and purposes of sanctions may pose limits on interchangeability. A study of the public’s sensibilities about fines and imprisonment in Canada revealed that regardless of severity, fines were not seen as accomplishing the traditional sentencing goal of denunciation as effectively as imprisonment. For some circumstances, namely violent or sexual offences, members of the public saw the need to denounce the crime, and believed that this could be accomplished most effectively through imprisonment (Marinos 1997).

Over time, imprisonment has been constituted and coupled with the expressive purposes of punishment — the goals of denunciation and general deterrence. The courts, media, criminal justice agencies and others are active participants in constituting

---

The term "sensibilities" refers to the feelings, values, and sentiments which are associated with, and support attitudes about different sanctions, and what it means to punish.
"sanctions" and reproducing meanings and sensibilities about their usefulness. In examining a range of special commission and committee reports in Canada, for instance, it is apparent that there has been an increasing reliance on conceptualizing and interpreting imprisonment as being effective in expressing condemnation of criminal behaviour. Beginning with the Ouimet Committee Report in 1969, to the Federal Government Task Force Report in 1977, we see how imprisonment has been actively linked with denunciation and general deterrence to symbolize the condemnation of 'criminals', while non-institutional sanctions have been linked to providing opportunities for the 'socially responsible' offender (see Ekstedt and Griffiths 1988: 56; Ouimet Committee 1969; Law Reform Commission 1975; MacGuigan Subcommittee 1977; Federal Task Force 1977). However, as the rationale of rehabilitation lost its credence and academic doubt increased about achieving general deterrence through sentencing (Canadian Sentencing Commission Report 1987: 138; Doob, Marinos and Varma 1995: 65-83), decisions to impose imprisonment are likely predominately linked with the purpose of conveying a denunciatory message (Manson 1997b: 290-3).

Short sentences of custody, in particular, may be relied upon by the courts in communicating messages about crime and punishment. Manson (1997b) argues that the only purpose for imposing a short sentence of imprisonment is to express denunciation. Using mandatory 14-day custodial sentences for impaired driving as an example, he argues that "[t]he rationale for the short mandatory sentence is its expressive denunciatory message: "Impaired driving is a crime." As with all short sentences, any presumption of incarceration is not based on general deterrence" (Manson 1997b: 292). In a recent sentencing decision in Ontario, Judge David Cole argued that in a case of
robbery, and in the particular circumstances of three first-time, youthful adult offenders, a minimum short period of custody (90 days) was required in order to express denunciation. After this denunciatory component was accomplished, the judge sentenced the offenders to substantial hours of community service (R. v. Visanjí).

The study on public perceptions of fines and imprisonment (Marinos 1997), and the link between imprisonment and denunciation suggests that sanctions are multidimensional. The present analysis suggests that it is wrong to assume that sanctions are highly fluid, flexible ‘things’ which can easily be calibrated into equivalencies at sentencing (Tremblay 1988). David Garland (1990, 1991), for instance, argues that there are various dimensions to punishment — moral, political, economic and cultural — and to evaluate punishment in merely instrumental terms is both misguided and unproductive. In order to understand the various dimensions of punishment, it is critical that the broader social and cultural contexts are examined. In particular, the nature of the offence for which the offender is being punished seems to be a critical aspect of understanding the sensibilities about different sanctions. Thus there are limits to the fluidity and flexibility of sanctions and the extent to which they can be “equivalent” on all dimensions. ‘Intermediate’ sanctions may not be ‘intermediate’ on certain dimensions.

**Understanding the use of ‘intermediate’ sanctions in youth courts**

Understanding that there are both qualitative and quantitative dimensions to punishment has important implications for examining the use of ‘intermediate’ sanctions in Canadian youth courts. The concept of “intermediacy” suggests that sanctions can be placed along a single continuum of severity — from most to least severe. The model of
the continuum is useful in highlighting a range of individual sanctions, as well as shifting our attention away from imprisonment — away from notions such as “alternatives” or “non-custodial” sanctions. However, the continuum model is problematic in reinforcing the view that sanctions only differ on a dimension of severity (Doob and Marinos 1995; Marinos 1997). In fact research on perceptions of ‘intermediate’ sanctions has generally focused on comparing sanctions by their severity (Tremblay 1988; Petersilia and Deschenes 1994; Harlow et al. 1995).

The limitations of conceptualizing punishment as a single continuum of severity have important implications for our understanding of young offender dispositions. Statistics published by the Canadian Centre for Justice Statistics (hereafter CCJS) demonstrate that as the “most significant disposition”, ‘intermediate’ sanctions are infrequently imposed by youth courts across Canada (see, for example, CCJS, 1994-5, Table 8; see also CCJS, 1998-1999). In examining statistics for 1994-5 (CCJS, Table 8), it is apparent that custody (secure and open combined) and probation are heavily used: they are the “most significant dispositions” in 32.7% and 48.2% of cases respectively. All other sanctions — those which may be considered ‘intermediate’ — are imposed as the “most significant disposition” in only 19.2% of all cases given dispositions in youth courts across Canadian provinces (n= 73,969 cases). Hence from published statistics of the “most significant disposition”, we would conclude that custody and probation are both frequently imposed compared to ‘intermediate’ sanctions.

*While terms like "alternatives to imprisonment" or "non-custodial punishments" are useful in shifting attention away from imprisonment, they reinforce it as the reference point of punishment. It is also noteworthy to point out that it is quite common to hear people use the term "sentenced" as the way to describe imprisonment: if an offender is not "sentenced" then he or she did not receive imprisonment. The association between sentencing and imprisonment is another example of the prison-punishment nexus — the ways in which "punishment" is connected to imprisonment.*
But it is critical to point out that the conclusion of the apparent low use of 'intermediate' sanctions relates to the way in which youth court dispositions are presented. On the one hand, the published statistics on youth court dispositions are informative in outlining in an easily understandable manner the varying use of sentencing options by youth court judges. On the other hand, the published statistics obscure the extent to which 'intermediate' sanctions are imposed. In the published statistics by CCJS, the dispositions handed down in a "case" are typically categorized by the single "most significant" disposition. The range of sanctions available is ranked along a single continuum — from most severe to least severe in a specified invariant order. Thus if a young offender was sentenced to open custody, probation and a fine, a count would appear under open custody as it is defined by CCJS as the "most severe" of these three. In this same example, the result is that the published statistics underestimate the use of 'intermediate' sanctions such as a fine. All 'intermediate' sanctions that are combined with probation are counted under the rubric of the punishment deemed "most significant" — "probation". Thus the published statistics on youth court dispositions underrepresents the imposition of 'intermediate' sanctions as 19.2%.

A number of important questions arise, then, from the discussion of the multiple dimensions of punishment and the limitations of published statistics of young offender dispositions in Canada. How frequently are 'intermediate' sanctions imposed? For which offences are 'intermediate' sanctions used alone and in combination with imprisonment? What do 'intermediate' sanctions accomplish? What is the purpose of combining an 'intermediate' sanction with custody? Moyer's recent analysis (1996) of youth court dispositions addressed combinations; however, the findings were not linked to the
broader purposes or sensibilities about punishment. As past literature suggests, ‘intermediate’ sanctions, in general, may be seen by members of the public, and presumably by judges, as lacking the symbolic and expressive qualities that imprisonment possesses. This likely affects how often and in which circumstances an ‘intermediate’ sanction is combined with custody (Doob and Marinos 1995; Marinos 1997).

By examining combinations of dispositions, it is possible to assess the purposes of the use of custody and ‘intermediate’ sanctions. Research on young offender dispositions reveals concerns about the increasing use of short sentences of custody since the introduction of the Young Offenders Act (Doob 1992). In a study of six provinces, Doob revealed that youth courts increased the proportion of cases getting short sentences of three months or less and youth courts are apparently handing down fewer very long sentences (1992: 81-82). Unfortunately, the published statistics on young offender dispositions do not provide information on the offences for which these short sentences are imposed, nor whether short custodial sentences are handed down alone or in combination with other sanctions. In fact statistics for 1994-5 demonstrate that sentences of imprisonment of thirty days or less make up over one-quarter of custodial sentences imposed in youth courts (CCJS, Table 6, 1994-5; see also Markwart 1992). Doob (1992) concludes his analysis by suggesting that “[a] more detailed analysis and a clearer statement of purpose of dispositions under the present legislation are necessary in order to evaluate these trends” (emphasis added) (1992: 83). An examination of the use of ‘intermediate’ sanctions is not only important in and of itself, but it will likely shed light on the purpose of imposing them along with short custodial sentences.
Method

Since the focus of the present analysis was to understand the use of ‘intermediate’ sanctions for young offenders in Canada, it was critical to explore the combinations of dispositions handed down in youth courts. As mentioned earlier, Canadian Centre for Justice Statistics presently publishes statistics on the “most significant disposition” handed down for each case disposed of in youth courts. Each case is associated with a single disposition ordered from the most to the least serious disposition, starting with custody (secure followed by open), probation, fine, compensation, pay purchaser, compensation in kind, community service order, absolute discharge, and “other” (CCJS, 1994-5: Table 8). While this is useful for many purposes, it has the unfortunate effect of burying combinations of dispositions.

Data on young offender dispositions were obtained from CCJS in a form in which cases were aggregated according to whether they included each of the possible dispositions listed in section 20(1) of the Young Offenders Act. In addition, because of the importance of custody and the high frequency of short custodial sentences (Doob 1992; Markwart 1992), CCJS provided a special code on whether or not the case included a total (open and secure) custodial sentence of 1-30 days, 31-180 days, and 181 days and more. It was important to know the total length of custody because it related to the theory that the lack of denunciatory power of ‘intermediate’ sanctions could be compensated by being attached to a short term of imprisonment. Thus cases involving

---

5Note that total (open and secure) custodial sentences were actually coded as 1 to 29 days, and 30 days, but then were pooled as total length of custody from 1 to 30 days. We wish to thank Denyse Carriere and Glen Doherty for their work in providing the data on youth court dispositions for Canada, 1994-5, in a form which showed multiple dispositions and total length of custody.
custodial sentences were described as having a total custody length of 1 to 30 days, 31 to 180 days and 181 days and over — described as “short”, “medium” and “long” total custodial terms.

As noted earlier, the current literature on intermediacy in punishment defines ‘intermediate’ sanctions as those lying between prison and probation (Morris and Tonry 1990). For the purposes of this analysis, ‘intermediate’ sanctions were considered to be a fine, restitution, compensation, pay purchaser, compensation in kind, community services, and “other”. Thus custody, probation, detain for treatment, prohibition/seizure/forfeiture, absolute discharge, and suspend driver’s license were not included in the measure of an ‘intermediate’ sanction. The data also included the type of offence as described in Tables 3 and 8 in CCJS statistics (1994-5).

Findings

The use of ‘intermediate’ sanctions

In examining the use (alone and in combination) of all dispositions (n= 73,969 cases), probation is imposed most frequently, in 65.4% (48,381) of all cases and custody is imposed in 34.1% (25,214) of all cases.6 ‘Intermediate’ sanctions were imposed in 42.9% (31,714) of all convicted cases. The finding that some ‘intermediate’ sanction is imposed in over 40% of all dispositions is noteworthy. The published “most significant disposition” statistics by CCJS, in contrast, revealed that ‘intermediate’ sanctions were the “most severe” sentencing options imposed in only 15.8% of the cases (CCJS, 1994-5, Table 8). Thus an examination of combinations of dispositions is important in

---

6In all analyses that follow, I examined the 73,969 cases in Canada in 1994-5 where there was a finding of guilt.
understanding the actual overall use of ‘intermediate’ sanctions.

For a selected number of relatively high frequency violent, property, drug and administration of justice offences, Table 4.1 reveals that ‘intermediate’ sanctions are imposed predominately between 30% to 45% of cases. For example, the range of offences receiving ‘intermediate’ sanctions includes violent offences such as robbery (34.3%) and break and enter (45.4%) as well as theft under/possession stolen property (47.8%) and mischief to property (56.0%). In looking at offences against the administration of justice, it is noteworthy to point out the relatively low use of ‘intermediate’ sanctions for “escape custody/unlawfully at large” (7.5%) compared to “other administration justice offences” (largely “failure to comply with disposition” and “failure to appear”) (34.4%). In examining the range of ‘intermediate’ sanctions, it becomes clear that community service is the most popular, imposed in over two-thirds of cases in which at least one ‘intermediate’ sanction is used (67.23%). Fines, on the other hand, are imposed in only 19% of all cases receiving at least one ‘intermediate’ sanction.
Table 4.1

The Use of 'Intermediate' Sanctions, Alone and In Combination with Other Sanctions, for Selected Offences

<table>
<thead>
<tr>
<th>Most Significant Charge</th>
<th>Received 'Intermediate' Sanction</th>
<th>No 'Intermediate' Sanction</th>
<th>Total</th>
<th>(N)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aggr. Assault/Assault with weapon</td>
<td>41.8%</td>
<td>58.2%</td>
<td>100%</td>
<td>2180</td>
</tr>
<tr>
<td>Assault</td>
<td>41.7%</td>
<td>58.2%</td>
<td>100%</td>
<td>7311</td>
</tr>
<tr>
<td>Sexual Assault</td>
<td>22.6%</td>
<td>77.4%</td>
<td>100%</td>
<td>879</td>
</tr>
<tr>
<td>Robbery</td>
<td>34.3%</td>
<td>65.7%</td>
<td>100%</td>
<td>1332</td>
</tr>
<tr>
<td>Possession Weapon</td>
<td>36.8%</td>
<td>63.2%</td>
<td>100%</td>
<td>1303</td>
</tr>
<tr>
<td>Break &amp; Enter</td>
<td>45.4%</td>
<td>54.6%</td>
<td>100%</td>
<td>10196</td>
</tr>
<tr>
<td>Theft Over</td>
<td>42.3%</td>
<td>57.7%</td>
<td>100%</td>
<td>2524</td>
</tr>
<tr>
<td>Vehicle Theft</td>
<td>44.1%</td>
<td>55.9%</td>
<td>100%</td>
<td>1065</td>
</tr>
<tr>
<td>Theft Under/Poss. Stolen Property</td>
<td>47.8%</td>
<td>52.2%</td>
<td>100%</td>
<td>15894</td>
</tr>
<tr>
<td>Mischief to Property</td>
<td>56.0%</td>
<td>44.0%</td>
<td>100%</td>
<td>4055</td>
</tr>
<tr>
<td>Possession Drug/Narcotic</td>
<td>51.2%</td>
<td>48.8%</td>
<td>100%</td>
<td>2047</td>
</tr>
<tr>
<td>Trafficking Drug/Narcotic</td>
<td>63.9%</td>
<td>36.1%</td>
<td>100%</td>
<td>1098</td>
</tr>
<tr>
<td>Escape Custody &amp; Unlawful at Large</td>
<td>7.5%</td>
<td>92.5%</td>
<td>100%</td>
<td>1899</td>
</tr>
<tr>
<td>Other Admin. Justice Offences'</td>
<td>34.4%</td>
<td>65.6%</td>
<td>100%</td>
<td>14992</td>
</tr>
<tr>
<td>All other cases</td>
<td>49.2%</td>
<td>50.8%</td>
<td>100%</td>
<td>7194</td>
</tr>
<tr>
<td>Total</td>
<td>42.9%</td>
<td>57.1%</td>
<td>100%</td>
<td>73969</td>
</tr>
</tbody>
</table>

7 This category of offences — "other administration of justice offences" — is largely made up of the offences of "failure to comply with disposition" and "failure to appear". It also includes "breach of recognizance", "failure to comply with probation order", "offence against the administration of justice", "failure to comply with undertaking", and "contempt against youth court".
Table 4.2

The Use of 'Intermediate' Sanctions Alone and In Combination with Prison and Probation

<table>
<thead>
<tr>
<th>Disposition</th>
<th>Percentage</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>'Intermediate' only</td>
<td>14.2%</td>
<td>10 485</td>
</tr>
<tr>
<td>'Intermediate' sanction &amp; probation only</td>
<td>25.8%</td>
<td>19 105</td>
</tr>
<tr>
<td>'Intermediate' sanction &amp; custody Only</td>
<td>0.6%</td>
<td>432</td>
</tr>
<tr>
<td>'Intermediate' sanction &amp; custody and probation only</td>
<td>2.3%</td>
<td>1 692</td>
</tr>
<tr>
<td><strong>Subtotal: 'Intermediate' alone or combination with other sanction</strong></td>
<td>42.9%</td>
<td>31 714</td>
</tr>
<tr>
<td>No 'intermediate' sanction</td>
<td>57.1%</td>
<td>42 255</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>100%</td>
<td>73 969</td>
</tr>
</tbody>
</table>

Table 4.2 presents the extent to which custody and probation are used in comparison to 'intermediate' sanctions. The findings demonstrate that when 'intermediate' sanctions are imposed, they are typically attached to probation. The findings reveal that, standing on their own, there is a relatively low use of 'intermediate' sanctions as 'intermediate' — imposed without prison or probation in only 14.2% of the cases. It is also interesting to note that while various authors have pointed to the high use of custody imposed for young offenders (Leschied and Jaffe 1987: 425), it is relatively rare that an 'intermediate' sanction is attached to custody (0.6%).
Table 4.3
The Proportion of those Cases that Received a Short, Medium or Long Custodial Sentence That Also Received an ‘Intermediate’ Sanction

<table>
<thead>
<tr>
<th>Offences</th>
<th>Short Prison %</th>
<th>N</th>
<th>Medium Prison %</th>
<th>N</th>
<th>Long Prison %</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>&quot;Serious&quot; Offences</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Aggr. Assault/With Weapon</td>
<td>17.2</td>
<td>233</td>
<td>9.8</td>
<td>429</td>
<td>3.9</td>
<td>154</td>
</tr>
<tr>
<td>Assault</td>
<td>9.9</td>
<td>825</td>
<td>6.9</td>
<td>728</td>
<td>4.5</td>
<td>89</td>
</tr>
<tr>
<td>Sexual Assault</td>
<td>17.9</td>
<td>56</td>
<td>3.8</td>
<td>157</td>
<td>9.9</td>
<td>131</td>
</tr>
<tr>
<td>Robbery</td>
<td>20.0</td>
<td>135</td>
<td>12.6</td>
<td>341</td>
<td>10.8</td>
<td>369</td>
</tr>
<tr>
<td>Break &amp; Enter</td>
<td>20.5</td>
<td>987</td>
<td>12.1</td>
<td>2373</td>
<td>10.0</td>
<td>912</td>
</tr>
<tr>
<td>Possession Weapon</td>
<td>7.0</td>
<td>129</td>
<td>5.5</td>
<td>201</td>
<td>3.7</td>
<td>54</td>
</tr>
<tr>
<td>Trafficking Drug or Narcotic</td>
<td>17.0</td>
<td>106</td>
<td>11.5</td>
<td>183</td>
<td>4.9</td>
<td>41</td>
</tr>
<tr>
<td><strong>&quot;Medium Serious&quot; Offences</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Theft Over</td>
<td>13.4</td>
<td>284</td>
<td>9.8</td>
<td>634</td>
<td>8.0</td>
<td>200</td>
</tr>
<tr>
<td>Vehicle Theft</td>
<td>10.1</td>
<td>109</td>
<td>8.0</td>
<td>100</td>
<td>0.00</td>
<td>10</td>
</tr>
<tr>
<td><strong>&quot;Less Serious&quot; Offences</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Theft Under/Possession Stolen Prop.</td>
<td>8.8</td>
<td>1760</td>
<td>8.0</td>
<td>1861</td>
<td>9.2</td>
<td>314</td>
</tr>
<tr>
<td>Mischief</td>
<td>14.5</td>
<td>351</td>
<td>11.4</td>
<td>315</td>
<td>12.5</td>
<td>40</td>
</tr>
<tr>
<td>Possession Drug or Narcotic</td>
<td>6.5</td>
<td>169</td>
<td>9.0</td>
<td>67</td>
<td>5.6</td>
<td>18</td>
</tr>
<tr>
<td><strong>Offences Against Admin. Justice</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Escape Custody/Unlawful at Large</td>
<td>2.4</td>
<td>1078</td>
<td>3.9</td>
<td>559</td>
<td>4.8</td>
<td>62</td>
</tr>
<tr>
<td>Other Administration Justice Offences*</td>
<td>5.6</td>
<td>4515</td>
<td>3.8</td>
<td>2085</td>
<td>6.3</td>
<td>223</td>
</tr>
<tr>
<td><strong>All Other Offences</strong></td>
<td>10.2</td>
<td>689</td>
<td>7.3</td>
<td>855</td>
<td>4.3</td>
<td>282</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>8.8</td>
<td>11426</td>
<td>8.1</td>
<td>10888</td>
<td>8.2</td>
<td>2899</td>
</tr>
</tbody>
</table>

* N's refer to the number of people from which each percentage is based.

*See Note 5 regarding Table 4.1.
Short "denunciatory sentences" and 'intermediate' sanctions

For certain types of offences — violent, sexual, or serious property — denunciation is likely to be an important purpose in sentencing a young offender. However, for the less serious instances of these offences (e.g., where the role of the particular offender was minor or where the particular offence was relatively minor), denunciation may still be seen as being important, but a custodial sentence may be seen as inappropriately harsh. One way to resolve this conflict is to impose a very short custodial sentence (e.g., a total sentence of 30 days or less) and combine it with the "appropriate" punishment (e.g., a community service order). If this logic is correct, then in instances where denunciation is important, one would expect 'intermediate' sanctions to be more likely combined with short custodial sentences than with longer ones where, presumably, the custodial sentence is seen as being the appropriate punishment.

Table 4.3 demonstrates that for offences requiring expressions of condemnation — violent, sexual and serious property — one or more 'intermediate' sanctions are more likely to be combined with short custodial sentences rather than long. This finding is consistent with the theory of the lack of denunciatory power of 'intermediate' sanctions. In looking at the offence of aggravated assault/assault with a weapon in Table 4.3, we see that of those who received a short custodial sentence, the proportion that received an 'intermediate' sanction with "short" custody was 17.2%. This is greater than the proportion that received an 'intermediate' sanction along with a medium (9.8%) or long

---

9 R. v. H. (S.R.), for example, stated that secure custody may be imposed for reasons other than protection of the public, including factors such as general deterrence and "the expression of society's abhorrence of certain crimes". The Ontario Court of Appeal held that "...the abhorrence of society of certain crimes must be taken into account by the youth court both when deciding whether there should be a period of custody and when deciding whether that custody should be open or secure or a combination of both" (at 50-51). A number of adult cases also have linked denunciation as an important purpose of sentencing violent and sexual offences, R. v. M. (C.A.) and R. v. M. (G.).
(3.9%) custodial sentence. In contrast, the effect is not the same for minor offences such as theft under/possession of stolen property: for each custody length, the proportion receiving 'intermediate' sanctions with a short (8.8%), medium (8.0%), or long (9.2%) custodial sentence did not vary substantially.

Thus the findings shown in Table 4.3 reveal that for "serious" and "medium serious" offences receiving different custody lengths — aggravated assault/assault with a weapon, assault, sexual assault, robbery, break and enter, possession of a weapon, theft over $1,000, trafficking of drugs or narcotics, and vehicle theft — 'intermediate' sanctions were more likely to be attached to a short period of custody than long. In contrast, for less serious offences receiving custody such as possession of stolen property, theft under $1,000, mischief to property, and possession of drugs or narcotics, the likelihood of receiving an 'intermediate' sanction did not vary substantially with custody length. Administration of justice offences not only had a low overall use of 'intermediate' sanctions, but they also were equally likely to be attached to each length of a custodial sentence. It is important to point out that the range of selected offences accounts for 90.3% of the total sentenced cases in 1994/5 (n = 73969 cases, total) and 93% (n = 24214 cases, total) of those who received custody.

Discussion and Conclusions

The use of 'intermediate' sanctions for young offenders

The assumption that sanctions serve multiple purposes, and that they are not highly fluid, flexible entities appropriate for a variety of offences and offenders led to the present inquiry of combinations of dispositions in youth courts across Canada. Recent
literature (Department of Justice 1992: Annex 1; Morris and Tonry 1990; Petersilia and Deschenes 1994) and legislation to increase the use of 'intermediate' sanctions assumes that sanctions can be made equivalent across a range of offences as long as they are roughly similar in severity or punitiveness. In contrast, poststructuralist writings on penalty and penal reform conceptualize 'punishment' as dispersed, fluid and multidimensional (Foucault 1977; Garland 1985, 1990; McMahon 1993), suggesting that penal reform is far more complex than simply focusing on the severity of sanctions. The very term *penality*, in fact, acknowledges that 'punishment' refers to the "whole of the penal complex, including its sanctions, discourses and representations, rather than simply stressing institutional practices" (Garland 1985: x). Thus while one might expect that sanctions are easily interchangeable across a range of offences and offenders, the present analysis suggests that there are some limits to the fluidity and flexibility of (criminal justice) sanctions.

In order to understand these limits, the social and cultural contexts of crime and punishment must be addressed. The nature of the offence for which the offender is being punished, as well as modern sensibilities about what is 'appropriate' and 'civilized' are important elements of these contexts. David Garland highlights the role of modern sensibilities in structuring the practices of contemporary penal systems:

But penal measures will only be considered at all if they conform to our conceptions of what is emotionally tolerable. The matter-of-fact administration of most penal policy is possible because it relies upon measures which have already been deemed tolerable and the morality of which can be taken for granted. . . . I do not mean by this that governments and penal authorities always take care to search their consciences, or put morality before expediency in their use of punishments — clearly they do not. But political decisions are always taken against a background of mores and sensibilities which, normally at least, will set
limits to what will be tolerated by the public or implemented by the penal system’s personnel (1990: 214).

The analysis of dispositions reveals that youth courts across Canada seem to be using ‘intermediate’ sanctions quite differently for different kinds of offences. In general, the findings of the present analysis demonstrate that the nature of the offence and purposes of punishment are important considerations in understanding the use of ‘intermediate’ sanctions and imprisonment. The analysis of combinations of dispositions reveals that they are imposed across a range of high frequency violent, property, drug and administration of justice offences — between 22% to 63% of these cases (Table 4.1). However, it was expected that ‘intermediate’ sanctions would more likely be imposed for property offences than violent and sexual offences. Consistent with this expectation, there were differences found in the use of ‘intermediate’ sanctions among offences such as mischief to property (56.0%) and theft under/possession of stolen property (47.8%), compared to possession of a weapon (36.8%), robbery (34.3%) and sexual assault (22.6%) (see Table 4.1).

But some authors have suggested that part of the importance of the conceptualization of “intermediate sanctions” — and moving away from notions like “alternatives to imprisonment” or “non-custodial sanctions” — is that individual sanctions are imposed alone, as importance sanctions on their own. For example, Morris and Tonry argue that both fines and community service orders should be taken seriously as credible sanctions and used more frequently, “standing alone or as part of a punishment package” (1990: 11). While ‘intermediate’ sanctions — community service in particular — are imposed fairly frequently, they may not be seen by youth court judges
as being capable of accomplishing what they want to accomplish on their own. Rather, when 'intermediate' sanctions are imposed, they are predominately attached to probation (Table 4.2). In light of the finding that community service is the most commonly imposed 'intermediate' sanction, it is likely that, at least in part, probation provides a mechanism for supervising these young offenders.

**The lack of denunciatory power of 'intermediate' sanctions**

Thus far, the analysis has focused on the use of 'intermediate' sanctions across various offences. For some offences, 'intermediate' sanctions are not seen by youth court judges as appropriate, and in other circumstances, only if tied to a custodial sentence. As mentioned earlier, an examination of combinations of dispositions is particularly important in trying to understand the purpose of short periods of custody for young offenders. Some authors have highlighted the denunciatory and expressive functions of short periods of custody, while attributing incapacitative and/or rehabilitative purposes to long periods of incarceration for young offenders. Thus it becomes particularly interesting to inquire into those cases that received a short period of custody that also received an 'intermediate' sanction. Examining those cases of short custody, compared to long custody, reveal a variation in the frequency and circumstances in which 'intermediate' sanctions are attached to custodial sentences.

The findings in Table 4.3 are consistent with the theory that 'intermediate' sanctions lack denunciatory power and therefore must be tied to a short custodial sentence when denunciation is important. These results are also consistent with a broader

---

10 While there is a relationship between the instrumental and expressive functions of punishment, the purpose here is to highlight how functions are conceptualized within the literature.
assumption that sanctions serve various purposes according to the nature of the offence and offender, and that ‘intermediate’ sanctions may not be as effective in accomplishing denunciation as imprisonment (Doob and Marinos 1995; Marinos 1997). In the present context, there has been concern expressed by members of the public that sentences for violent offences are not severe enough and that there is a need for increased accountability of young offenders under the Young Offenders Act (Sprott 1996). One way to accomplish denunciation and accountability when custody, per se, is not seen as being justified is to impose an ‘intermediate’ sanction along with a short custodial sentence.

This finding is consistent with some literature on the denunciatory function of short periods of custody. Manson’s analysis of conditional sentences of imprisonment — an ‘intermediate’ sanction for adult offenders — highlights the denunciatory purpose of imprisonment. As mentioned earlier, he argues that the only reason for imposing imprisonment, and short periods in particular, is to denounce crime (Manson 1997b). Attaching an ‘intermediate’ sanction to a short term of custody, then, would allow judges to fulfill the denunciatory component required for some offences and therefore justify or enable the use of the ‘intermediate’ sanction.

There is evidence to suggest that since the introduction of the Young Offenders Act, there has been a trend towards short custodial sentences, and has been explained as the “short sharp shock” model of sentencing (Doob 1992: 24; Markwart 1992: 267; Doob et al. 1995: 139). This observation might be used to explain the pattern of results presented in Table 4.3. In the present analysis, “short sharp shock” does not adequately explain custodial sentences combined with other sanctions. It might be possible to
explain the variation in the frequency of custodial sentences and ‘intermediate’ sanctions by “just proportionality”. The present findings would suggest that in fulfilling the minimum denunciatory component of the sentence through short custody, judges are then able to impose ‘intermediate’ sanctions. However, both theories would not have led to the present inquiry into combinations of dispositions, and in particular, different custody lengths combined with an ‘intermediate’ sanction across some offences and not for others. “Short sharp shock” and proportionality would not lead one to predict the finding that for “serious” and “medium serious” offences receiving different custody lengths — compared to “less serious” offences — ‘intermediate’ sanctions were more likely to be attached to a short period of custody than long.

The finding of the lack of denunciatory power of ‘intermediate’ sanctions combined with the high use of short sentences of custody suggests that we need to find new ways to accomplish this goal and more broadly, assert community values. It may be instructive in the context of sentencing young offenders to learn from some of the present problems in the adult sentencing process — the development of the conditional sentence of imprisonment. It may be that conditional sentences of imprisonment will be less frequently imposed in circumstances that require denunciatory sentences by the court — namely for violent and sexual offences. In fact, courts across Canada have been grappling with how to accomplish denunciation through conditional sentences of imprisonment.

As stated earlier, it seems that imprisonment has been coupled with the expressive functions of punishment, making it difficult for other sanctions to be seen as capable of accomplishing these goals. Recently, the Ontario Court of Appeal in Wismayer (1997) attempted to uncouple the link between imprisonment and denunciation through the
imposition of a conditional sentence of imprisonment for a sexual assault. Justice Rosenberg argued that “I cannot accept that a conditional sentence of imprisonment is unavailable where the paramount consideration is denunciation of the offender’s conduct” (1997: 33). In this case for the Ontario Court of Appeal, denunciation was accomplished through the imposition of strict conditions of house arrest, monitored by the offender’s parents. It may be possible, then, to decrease the use of some sentences of short custody for young offenders through more creative conditions of community supervision (such as house arrest). While efforts must be made in the courts to uncouple the nexus between imprisonment and denunciation, it is also necessary to acknowledge the limits of our sensibilities. It seems that we value autonomy and security of the person to the extent that only imprisonment, in some circumstances, meets this expectation of punishing offenders.

The findings may explain the increased use of short sentences of custody since the introduction of the Young Offenders Act (Doob 1992). It is possible that judges are reacting to public disapproval about youth crime and young offenders' sentences, and this might explain the increase in short sentences of custody for young offenders. As the public demanded “denunciation” and strong statements about crime through sentencing, this was being accomplished increasingly by combining a short custodial sentence with an ‘intermediate’ sanction.

In the (adult) case referred to earlier (Visanji (1997)), Judge Cole explained the purpose of short custody — ninety days — combined with an ‘intermediate’ sanction for three first-time youthful adult offenders convicted of robbery:
I have decided that it would be consistent with Parliament's direction that judges are to reduce the use of incarceration, expressed in ss. 718.2(d) and (e) of the Code, to impose community service as a substitute for additional imprisonment, once the minimum denunciatory period of 90 days incarceration has been satisfied (emphasis added) (at 417).

In the case noted above, the judge found that a conditional sentence of imprisonment — even with the mandatory and optional conditions which could be imposed — would be "insufficient to meet the principles of denunciation" (23). The "appropriate minimum" term of custody was required to fulfill the denunciatory portion of the sentence for robbery, which then allowed the judge to impose community service as a substitute for additional incarceration as well as restitution to the victim. It is also interesting to note that the judge saw this particular robbery — committed by first-time youthful adult offenders causing some physical and psychological harm to the victim — as being appropriate for an 'intermediate' sanction, but also requiring a short denunciatory sentence of custody.

Understanding the combination of custody and 'intermediate' sanctions is important because it may help explain the relatively frequent use of custody for relatively minor offences. As provisions under the Young Offenders Act specifically relate to imposing custody only when necessary, it is critical to inquire into the variations in the use of 'intermediate' sanctions and the purposes of different sanctions. A focus on the combination of custody and 'intermediate' sanctions is useful in exposing the qualities and complexities of punishment. Furthermore, when one considers that 15.4% of all dispositions handed down in 1994/5 involved a period of custody of 30 days or less, and that short dispositions constituted 45.3% of all custodial dispositions handed down to young offenders in 1994/5, it is important to consider whether there might be a more
sensible way of accomplishing the relevant goals of sentencing.

Whether an 'intermediate' sanction is seen as appropriate varies according to the nature of the offence and offender. In the present context of young offenders in Canadian courts, the use of 'intermediate' sanctions does not seem to be related to what is seen as an 'intermediate' crime. The nature of the offence and what is seen as important to be accomplished at, and through the sentencing process are likely important considerations for youth court judges.
Chapter 5
Judicial Sensibilities Toward Conditional Sentences of Imprisonment
and the Denunciatory Function of Punishment

Introduction

Canada’s most explicit form of interchangeability is the conditional sentence of imprisonment. Developed as a replacement for and alternative to imprisonment of less than two years (section 742.1 Criminal Code), the conditional sentence is intended to be equivalent to prison. Like Morris’s and Tonry’s (1990) proposal for interchangeability, however, Parliament included few guidelines on the use of conditional sentences, and assumed that the sanction would be appropriate for a wide range of offences and sentencing purposes. Thus trial court judges enjoy wide discretion in the imposition of conditional sentences.

However the conditional sentence has posed various problems for courts at both the trial and appeal court levels across the provinces. One area of concern relates to the role of the public. Cases such as Brady, Parker, W.L.F., and Collins1 have highlighted the role of public attitudes towards the use of conditional sentences. Justice Rosenberg has argued that in light of pronouncements by the courts -- using Parker as an example -- it is important that the public do not perceive the courts as using conditional sentencing for inappropriate cases because support for this sanction and the administration of justice will be undermined (Rosenberg 1999: 61). More broadly, these cases follow the spirit of the Supreme Court of Canada’s decision in C.A.M. where it was stated that sentencing

reinforces community values. Thus decisions about the appropriate use of conditional sentences not only are seen by the courts as having an impact on the offender, but also on the wider community in which sentencing takes place.

Second, a number of courts have pointed to the perceived and operational lack of equivalence between the conditional sentence and imprisonment. There has been considerable debate in the media, the courts, and among researchers about the extent to which the conditional sentence is different from a suspended sentence with probation (Rosborough 1998; Gemell 1997; Brady; Roberts, Antonowicz and Sanders 2000), and whether the conditional sentence lacks restrictive conditions compared to the traditional prison. As a result, the courts have grappled with whether the conditional sentence is qualitatively different from probation and whether it is sufficiently equivalent to the traditional prison (see Brady; Smith). For example, in Smith the Ontario Court of Appeal described the need for strict conditions as part of a conditional sentence:

It [the conditional sentence] is intended to achieve the purposes of a custodial sentence without institutional incarceration. In order to achieve those purposes there should generally be some restrictions on the liberty of the person during the period of the sentence. This usually takes the form of some variation of house arrest... (at paragraph 11, 1999).

The third area relates to the appropriate use of conditional sentences compared to imprisonment for violent and sexual offences (see Manson 1997a). Sexual assault cases

---

have created considerable difficulties and controversy for courts, particularly those involving breach of trust (i.e. B. (S.); Matheson; Nikkanen). Intimately related to the nature of the offence is whether the conditional sentence can be seen as accomplishing the purposes of denunciation and general deterrence as effectively as prison in cases involving violence (i.e., Wismayer). In fact, questions about the appropriateness of conditional sentences for sexual offences and the ability of this punishment to accomplish denunciation as effectively as imprisonment were recently heard by the Supreme Court of Canada in a series of appeals. The Supreme Court’s decision of Proulx stated that a conditional sentence can achieve denunciation and general deterrence – the expressive function of punishment – as effectively as imprisonment with the help of restrictive conditions such as house arrest and electronic monitoring. Clearly judges can face difficult issues when deciding to impose a conditional sentence of imprisonment.

The sentencing purpose of denunciation is of critical importance in the discussion of conditional sentences and is therefore central to this Chapter. Chapter 3 provided evidence that members of the public do not fully support conditional sentences for violent offences because of its perceived inability to accomplish denunciation. Chapter 4 revealed that in examining dispositions, youth courts are less likely to impose an ‘intermediate’ sanction alone for violence. Rather, the courts combine of a short period of custody with

---

an 'intermediate' sanction in order to fulfill the denunciatory component of the sentence.

In both of these research sites, non-prison sanctions were viewed as relatively inadequate to accomplish denunciation as effectively for violence compared to prison. The present Chapter examines judges' views about achieving denunciation with a conditional sentence of imprisonment. It then considers whether this factor might help explain other attitudes they have about this relatively new sanction. Examining the role of denunciation in the context of conditional sentences becomes an interesting way to understand judges' sensibilities about the notion of substituting sanctions for imprisonment and the multiple dimensions of punishment.

There has been limited research conducted on judges’ attitudes and values towards sentencing in Canada. Doob has argued for the importance of sentencing principles in guiding judicial decision-making and reducing disparity in the context of the Young Offenders Act and adult offenders (Doob 1989: 203; Doob 1990). Second, John Hogarth's Sentencing as a Human Process (1971) found that judges' penal philosophies directly related to their decision-making. Unfortunately, denunciation was not included in Hogarth's study as a purpose in sentencing offenders in this study.12


---

11 The term 'intermediate' sanctions is found within italics in order to signify that it is socially constructed (see Peterson, R. D. and D. J. Palumbo. 1997. “The Social Construction of Intermediate Punishments.” The Prison Journal. 77(1): 77-91). I use the term this way in order to question the very notion that what is "intermediate" between probation and probation varies according to the quantitative and qualitative dimensions of punishment -- the nature of the sanction, the offence, the purpose of the sanction, and the age of the offender.
appropriate responses to unwarranted disparity, and different models of guidance in sentencing decisions. The majority of respondents (88%) supported protection of the public as the overall purpose. However the purpose of denunciation was not considered in the questions posed to judges. Finally a study of sheriffs' views about the imposition of fines in Scotland focuses upon the objectives of sentencing (Young 1989). Young revealed that the fine, compared to other sentencing options, was seen by sheriffs to be the most "straightforwardly justifiable in retributivist/punishment terms" (1989: 47-49). Thus there has been relatively little inquiry into judges' sensibilities -- their attitudes, values, and meanings -- about the purposes of sentencing and the imposition of sanctions, and in particular substituting a sanction for a sentence of imprisonment.

At the institutional level, there has been scant public discussion about the conditional sentence. In fact Judge David Cole has remarked that "the legislation was passed without proper debate and judges were inadequately prepared to begin interpreting it." (Makin 2000). Clearly an examination of judges' attitudes towards the conditional sentence is critical if it will be a true substitute for traditional prison sentences of less than two years.

More specifically, it is important to inquire, to what extent are judges viewing the conditional sentence as interchangeable with imprisonment across different purposes and

---

12 Hogarth's analysis included the "classical purposes" of reformation, general deterrence, individual deterrence, incapacitation, and punishment (Hogarth 1971: 70).
13 The sheriff is described by Young as: "[S]heriffs are not lay judges; rather they are members of the legal profession who have at least ten years' experience as solicitors or advocates. Sheriffs sit alone in their courts" (1989: 47).
14 However a preliminary inquiry into judges' views about conditional sentences involved interviews with eleven judges in three local courthouses in the Toronto area. See LaPrairie, C. and C. Koegl. 1998. "Conditional Sentencing: Exploratory research into opinions and perceptions of judges in select urban courts." Department of Justice Canada.
offences? To what extent do judges see denunciation as being accomplished through the use of conditional sentences? Do these views about the ability to achieve denunciation through the conditional sentence explain other attitudes that judges have about this sanction, compared to imprisonment? In particular, do their views about denunciation relate to their attitudes about the public's support and understanding of the sanction? Does denunciation relate to their views about the differences between probation, a conditional sentence, and prison? Finally, does this purpose explain judges' sensibilities about the suitability of violent offences for the conditional sentence compared to prison?

The purpose of the present analysis is to identify and present patterns of responses.

In the present Chapter, I argue that denunciation plays an important role in judges' sensibilities about the conditional sentence compared to the traditional prison. I suggest that judges will see this purpose as being difficult to achieve through the use of conditional sentences. It is likely that their views about the ability of the conditional sentence to achieve this goal will relate to their attitudes about the lack of public understanding and support for the sanction, judges' perceptions about the similarity between probation and a conditional sentence, and perceptions about the unsuitability of a conditional sentence for violent and sexual offences. Denunciation likely will explain why some judges believe that the sanction should further restrict an offender's liberty. I argue that on the whole, judges will see the conditional sentence as less easily interchangeable than the legislation assumes. It will not be viewed as interchangeable across all purposes and offences, and will be viewed more like a sanction within the community than a sentence of imprisonment.

& Mail, A6.
Methods

In May 1998, provincial court judges across the country were asked to fill out a survey about their views on conditional sentences of imprisonment. Four hundred and sixty-one trial court judges responded, representing 36% of the total population. This is a respectable response rate for a busy professional sample. The survey was carried out by the Department of Justice, and a final report was produced in the Spring 1999.

In addition to trial court judges’ views, some interviews were conducted with Court of Appeal judges from two provinces. The interviews are also focused upon the relationship between denunciation and other views held by judges on conditional sentences. To what extent do Court of Appeal judges’ attitudes about the denunciatory aspects of conditional sentences relate to their other views about this sanction? Like the survey of trial court judges, interviews with Court of Appeal judges also focused upon issues about their perceptions of achieving denunciation through a conditional sentence, public understanding and support of the sanction, conceptualizations of the conditional sentence compared to probation and traditional prison, and judges’ views about the

---

16 "...[T]he response rate compares favourably with other criminal justice surveys. The last systematic survey of sentencing judges across Canada was conducted in 1986, and generated a response rate of 32% (see Research Staff of the Canadian Sentencing Commission, 1988)" in Roberts, Doob and Marinos. 1999. Judicial Attitudes to Conditional Terms of Imprisonment: Results of a National Survey. Report to the Department of Justice Canada.

17 A final report was produced by Roberts, J.V., Doob, A.N., and V. Marinos (with the assistance of Carol LaPrairie, Judge David Cole, and Tracy Perry). 1999. "Judicial Attitudes to Conditional Terms of Imprisonment: Results of a National Survey." Report to the Department of Justice Canada. The following Chapter is based upon the survey data, but the guiding research questions and analyses differ from the report.

18 One Justice in a third province (which would make a total of 8) was asked about being interviewed on his/her views about conditional sentences of imprisonment. The Justice responded that there was “no point” in an interview about views because he/she simply applies the law and thus responded that “my views don’t matter.”
appropriateness of the conditional sentence for violent and sexual offences.

While an examination of a large sample of interviews with Appeal Court judges on conditional sentences would have been ideal, it was difficult to gain access to them. However some interviews were conducted in person and over the telephone. For two provincial Courts of Appeal, a letter was sent to the Chief Justices, and then to various judges explaining the purpose of the interview. All judges then were contacted by telephone on a particular date in order to gain consent.

FINDINGS

Judicial Sensibilities Toward Interchangeability of Conditional Sentences for Sentences of Imprisonment

Are conditional sentences viewed as interchangeable with imprisonment across sentencing purposes?

Since the conditional sentence essentially replaces a custodial term served in a traditional prison, it is important to understand whether judges view the sanction as accomplishing denunciation as effectively as imprisonment.19 If conditional sentences and traditional prison sentences of less than two years are viewed as being similar and, therefore, interchangeable across sentencing purposes, then one would expect that judges would see the two as accomplishing denunciation equally.

All respondents (461) were asked whether they were “able to set conditions for a conditional sentence that are as effective as a normal sentence of imprisonment in accomplishing each of the following principles or purposes of sentencing”. This included

---

19 For example, see Wismayer (1997) 115 C.C.C. 3d 18 (Ont.C.A.).
“ensuring that the sentence is proportionate to the gravity of the offence and the degree of responsibility of the offender (proportionality); denouncing unlawful conduct (denunciation); deterring the offender and other persons from committing offences (deterrence); assisting in the rehabilitation of offenders (rehabilitation); and providing reparations for harm done to the victim or to the community (victim reparation)”.

Respondents answered whether a conditional sentence can accomplish each purpose on a five-point scale, including “always”, “usually”, “sometimes”, “almost never”, or “never”. Table 5.1 presents results of trial court judges’ responses about the ability of the conditional sentence to accomplish each sentencing purpose as effectively as imprisonment.

**Table 5.1**

**Respondents’ Views About a Conditional Sentence Accomplishing the Purposes of Sentencing as Effectively as Imprisonment**

<table>
<thead>
<tr>
<th>It can be as effective...</th>
<th>Proportionality</th>
<th>Denunciation</th>
<th>Deterrence</th>
<th>Rehabilitation</th>
<th>Reparation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Always/ Usually</td>
<td>51.1%</td>
<td>35.3%</td>
<td>34.7%</td>
<td>71.7%</td>
<td>59.2%</td>
</tr>
<tr>
<td>Sometimes</td>
<td>33.6%</td>
<td>33.0%</td>
<td>40.9%</td>
<td>23.8%</td>
<td>30.7%</td>
</tr>
<tr>
<td>Almost Never/Never</td>
<td>15.3%</td>
<td>31.7%</td>
<td>24.4%</td>
<td>4.4%</td>
<td>10.1%</td>
</tr>
<tr>
<td>Total</td>
<td>100% (450)</td>
<td>100% (448)</td>
<td>100% (450)</td>
<td>100% (449)</td>
<td>100% (443)</td>
</tr>
<tr>
<td>Mean*</td>
<td>2.80</td>
<td>3.28</td>
<td>3.14</td>
<td>2.37</td>
<td>2.61</td>
</tr>
</tbody>
</table>

*Low numbers indicate more effectiveness in accomplishing the goal.

Almost three-quarters (72%) of trial court judges who were surveyed believed that
the conditional sentence was "always" or "usually" as effective as prison in achieving rehabilitation. More than half of the respondents also viewed the conditional sentence as effective in accomplishing victim reparation (59.2%) and proportionality (51.1%). In contrast to the results found for rehabilitation, however, just over one-third (35.3%) believed that this was true for denunciation. A comparison of the mean differences on the original five-point scale was conducted to understand the variability between judges' ratings of rehabilitation and denunciation. The mean differences are significantly different, confirming that judges made distinctions between the two purposes. Since rehabilitation is most often associated with sanctions in the community, the results might suggest that judges view the conditional sentence more like a sanction within the community, rather than equivalent to imprisonment.

Like trial court judges, Court of Appeal judges also suggested that they see the conditional sentence of imprisonment as lacking denunciatory power, compared to imprisonment, and that they believe it has strong rehabilitative value. The denunciation-rehabilitation dichotomy is apparent in their responses to three questions: "What do you think is accomplished through the development of a conditional sentence in terms of the goals of sentencing (denunciation; general deterrence; individual deterrence; assist rehabilitation; victim reparation; accountability of offenders)?" "Do you think that a conditional sentence of imprisonment can accomplish denunciation and general deterrence as effectively in prison? Why or why not?" and "Do you think that imprisonment has traditionally been seen by judges as the way to accomplish

---

20 t = 18.96, df= 445, p<.001
denunciation or general deterrence?"21

For the first question, some respondents (3 out of 7) pointed to the benefits of pursuing rehabilitation through the use of the conditional sentence. For example, Judge #4 stated that "The biggest goal is the rehabilitative part -- working more on the offender rather than the offence." Another judge responded: "...With the conditional sentence we can better achieve rehabilitation, better integration in society. There are better structured drug programs, programs for repeat offenders etc." (Judge #3).

Five out of seven respondents specifically mentioned denunciation as a specific goal that, to some extent, poses a challenge to accomplish with the use of a conditional sentence. For instance, Judge #2 noted the link between the inability to achieve denunciation with this sanction and the role of the public:

...The conditional sentence does not accomplish general deterrence or denunciation; any factors to do with public perception of the role of sentencing is completely undermined by the conditional sentence.

One judge (#7) illustrates the denunciation-rehabilitation dichotomy in terms of achieving rehabilitation predominately through the community, and ensuring the 'other' is done -- "justice" or denunciation -- through imprisonment:

Perhaps in some cases, we can accomplish all of these goals with a conditional sentence. If you have a relatively young offender, he or she could be reeducated, turn them around, their thoughts and attitudes. Where the public get cynical is where you have one with a lengthy record or violent offence and they are out in the community. Then the sanction is viewed as a joke, and the administration of justice is viewed as a joke. An important element which is often overlooked is public confidence in the administration of justice. There must be a balance.

---

21 The interview questions are included in an appendix to this Chapter.
Judge #1 stated that the conditional sentence

...keeps offenders in the community and keeps their roots in the community. The problem with prison is that we yank them out of the community. The conditional sentence allows them to stay in the community and keep them employed and out of crime. Many offenders have psychiatric problems or addictions. We know that treatment programs are less effective in prison and better in the community.

One judge responded critically when asked “Do you think that a conditional sentence of imprisonment can accomplish denunciation and general deterrence as effectively as imprisonment? Why or why not?” (original emphasis). Judge #2 stated:

Denunciation is an exaggerated feature of sentencing. It plays into our assumption that judges are above everyone. We forget how denunciatory sentencing is, how embarrassing the very process is to your employer, friends, family. So to ask the question is to answer it. It is an attempt at euphemism. We’re dreaming. You cannot expect the public to think that being in the comfort of your home, reading, and watching t.v. is like jail. It is a joke; not the same loss of liberty. Conditional sentences will cause the public to lose integrity in the system, and integrity in the system is important.

Finally, when asked about whether imprisonment has traditionally been seen as the way to accomplish denunciation and general deterrence -- the expressive goals of sentencing -- all judges agreed. They responded that these goals have been most often associated with imprisonment. All seven Court of Appeal judges who were interviewed saw the connection between imprisonment on the one hand, and denunciation and general deterrence on the other hand as problematic. They described the implications of this connection as posing limits on their ability to find creative alternatives to imprisonment, and creating a cultural preoccupation with prison. Three examples are provided:

Yes. That’s what’s wrong with our system. They [the courts and the public] assume that prison is the only means. That’s why we rely too much on prison and
are not very imaginative. Scandinavian countries use other ways other than prison to accomplish these goals (Judge #1).

Yes. Prison is depended upon and it stems from the very structure of the Canadian Criminal Code. The Code is structured in a way that makes prison the focal point. You see the maximum penalties for offences. It drives people to thinking of penalties as jail...The way the Code is structured is really a factor. It easily accommodates the thinking (Judge #6).

I think so. But judges are reflecting society's values; judges are also part of society and are trying to exercise society's psyche -- the same thing. The best way to denounce is to send to jail (Judge #4).

The responses by Court of Appeal judges reinforce trial court judges' views of the ability of the conditional sentence to accomplish denunciation and rehabilitation. It is clear that trial court and Appeal Court judges identify the conditional sentence as being more effective as a sanction to achieve rehabilitation than denunciation. The prison is seen as a means of expressing condemnation of criminal behaviour, and acceptability of the sanction to the public seems important in this process. Clearly judges do not see the conditional sentence and imprisonment as interchangeable or equivalent when thinking about the purposes of sentencing.

*Are Judges' Views About the Ability to Achieve Denunciation With the Conditional Sentence Related to Other Beliefs They Hold About this Sanction?*

*Judicial Perceptions About Public Attitudes Towards Conditional Sentences*

Public acceptability is critical with the development of a new sanction that has as its aims to replace and reduce the use of imprisonment. In fact the Supreme Court of Canada in *C.A.M.* -- a sentencing appeal not involving a conditional sentence -- expressed
the importance of public attitudes toward the administration of justice (R. v. M.(C.A.)). The case acknowledged that denunciation is a legitimate and important objective of sentencing, sending a message both about the abhorrence of the offender's conduct and reinforcing community values.

Some comments by Court of Appeal judges also stressed the role of the public as the receiver of denunciatory messages through sentencing practices. If judges conceptualize interchangeability of a conditional sentence and imprisonment falling along a single dimension -- that only the severity of the offence and sanction are important -- then they would likely perceive that the public fully understands and supports the substitution of these sanctions. But if judges believe that punishment involves multiple dimensions beyond severity -- interchangeability might not be appropriate across all purposes -- then their perceptions about the public might be more complex. It is likely that judges who are critical about the ability to achieve denunciation through the conditional sentence -- and thus are not interchangeable across sentencing purposes -- would be skeptical about the extent to which the public fully understands and supports this sanction compared to imprisonment.

A number of questions about the public's perceptions of conditional sentences were asked in the survey of trial court judges. For example, one question asked, "Do you feel the general public understands the nature of conditional sentences?" Another question followed: "Do you feel the members of the general public who are aware of the nature of the conditional sentences support their use?" (original emphasis). The answers were given on a 6-point scale and pooled to 4 responses as "yes, all and most of the public", "yes, some of the public", "only few and none of the public", and "don't know".
Tables 5.2 and 5.3 present the results of the two questions.

### Table 5.2

**Judges’ Views About the Effectiveness of a Conditional Sentence in Achieving Denunciation and Their Views About Public Understanding of the Sanction**

<table>
<thead>
<tr>
<th>Denunciation is Achieved by a Conditional Sentence...</th>
<th>Most &amp; Some of the Public*</th>
<th>Few &amp; None of the public</th>
<th>Don't know</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Always/Usually</td>
<td>25.9% (41)</td>
<td>59.5% (94)</td>
<td>14.6% (23)</td>
<td>100%</td>
</tr>
<tr>
<td>Sometimes</td>
<td>15.1% (22)</td>
<td>71.9% (105)</td>
<td>13.0% (19)</td>
<td>100%</td>
</tr>
<tr>
<td>Almost Never/Never</td>
<td>9.2% (13)</td>
<td>50.4% (71)</td>
<td>40.4% (57)</td>
<td>100%</td>
</tr>
</tbody>
</table>

*No judge responded that the public understands the nature of the conditional sentence of imprisonment “all of the time”.

χ² = 49.016, df= 4, p<.0001

Table 5.2 illustrates that few judges (9.2%) who do not believe that a conditional sentence is effectively denunciatory (“almost never” and “never”) perceive that the public understands the nature of the sanction “most” or “some of the time”. In contrast, 25.9% of those judges who see the sanction as effectively denunciatory (“always” and “usually”) have confidence that the public understands the nature of the conditional sentence. Looking at Table 5.3 below, a similar relationship was found when judges were asked about public assent of the sanction (by those who are aware of its nature). Those judges who “almost never or never” believed that denunciation could be achieved with a conditional sentence were significantly less likely to believe that members of the public (who understand its nature) support its use.
Continuing with the theme of judges' considerations about the public, the survey asked respondents "Do you ever consider the impact that a conditional sentence order might have on public opinion?" It is not surprising that judges do not feel they need to consider public opinion in light of the finding in Table 5.3. Judges who perceived that denunciation can be achieved also believed that most 'aware' members of the public support the use of the conditional sentence. The relationship between judges’ views of denunciation and their consideration of the impact of a conditional sentence on public opinion is presented in Table 5.4 below.
Table 5.4

Judges' Views About the Ability of a Conditional Sentence to Accomplish Denunciation and Their Consideration of the Impact of a Conditional Sentence on Public Opinion

<table>
<thead>
<tr>
<th>Denunciation in Achieved by a Conditional Sentence</th>
<th>All of the Time</th>
<th>Most/Some of the Time</th>
<th>No, Never</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Always/Usually</td>
<td>16.1% (25)</td>
<td>58.1% (90)</td>
<td>25.8% (40)</td>
<td>100% (155)</td>
</tr>
<tr>
<td>Sometimes</td>
<td>16.0% (23)</td>
<td>68.1% (98)</td>
<td>16.0% (23)</td>
<td>100% (144)</td>
</tr>
<tr>
<td>Almost Never/Never</td>
<td>24.3% (33)</td>
<td>58.8% (80)</td>
<td>16.9% (23)</td>
<td>100% (136)</td>
</tr>
</tbody>
</table>

$\chi^2 = 9.313$, df $= 4$, p < .05

In order to test the importance of denunciation in explaining judicial views about public acceptability of the conditional sentence, it was necessary to explore the same relationships with their beliefs about rehabilitation. Since the hypothesis is focused on the purpose of denunciation exclusively, one would expect that judges' views about rehabilitation would result in a different pattern. Tables 5.5, 5.6, and 5.7 present results of judges' views on achieving rehabilitation through a conditional sentence and its relationship to their perceptions about the public's understanding and support of this sanction, and considerations of the impact of the sanction on the public.
Table 5.5
Judges’ Views About the Effectiveness of a Conditional Sentence in Achieving Rehabilitation and Their Views About Public Understanding of The Sanction

...General public understands the nature of the conditional sentence

<table>
<thead>
<tr>
<th>Rehabilitation is Achieved by a Conditional Sentence...</th>
<th>Most &amp; Some of the Public*</th>
<th>Few &amp; None of the public</th>
<th>Don’t know</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Always/ Usually</td>
<td>19.9% (64)</td>
<td>64.5% (207)</td>
<td>15.6% (50)</td>
<td>100% (321)</td>
</tr>
<tr>
<td>Sometimes</td>
<td>8.6% (9)</td>
<td>56.2% (59)</td>
<td>35.2% (37)</td>
<td>100% (105)</td>
</tr>
<tr>
<td>Almost</td>
<td>15.0% (3)</td>
<td>30.0% (6)</td>
<td>55.0% (11)</td>
<td>100% (20)</td>
</tr>
<tr>
<td>Never/Never</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*No judge responded that the public understands the nature of the conditional sentence of imprisonment “all of the time”.

\[\chi^2 = 34.553, \text{df}= 4, p<.001\]

Table 5.6
Judges’ Views About the Effectiveness of a Conditional Sentence in Achieving Rehabilitation and Their Views About Public Support for Use of Sanction

...Public who are aware of the nature of conditional sentence support their use

<table>
<thead>
<tr>
<th>Rehabilitation is Achieved by a Conditional Sentence...</th>
<th>All, Most &amp; Some of the public who are aware</th>
<th>Few &amp; None of the public who are aware</th>
<th>Don’t know</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Always/ Usually</td>
<td>63.4% (201)</td>
<td>23.0% (73)</td>
<td>13.6% (43)</td>
<td>100% (317)</td>
</tr>
<tr>
<td>Sometimes</td>
<td>34.0% (36)</td>
<td>50.0% (53)</td>
<td>16.0% (17)</td>
<td>100% (106)</td>
</tr>
<tr>
<td>Almost</td>
<td>15.8% (3)</td>
<td>73.7% (14)</td>
<td>10.5% (2)</td>
<td>100% (19)</td>
</tr>
<tr>
<td>Never/Never</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\[\chi^2 = 47.932, \text{df}= 4, p<.0001\]
Tables 5.5 & 5.6 show similar findings to those relating to denunciation. Judges who thought that rehabilitation could be accomplished were more likely to believe that the public understand the nature of the conditional sentence and support its use.

Table 5.7

Judges' Views About the Ability of a Conditional Sentence to Accomplish Rehabilitation and Their Consideration of The Impact of a Conditional Sentence on Public Opinion

<table>
<thead>
<tr>
<th>Rehabilitation in Achieved by a Conditional Sentence</th>
<th>All of the Time</th>
<th>Most/Some of the Time</th>
<th>No, Never</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Always/ Usually</td>
<td>15.3% (48)</td>
<td>64.3% (202)</td>
<td>20.4% (64)</td>
<td>100% (314)</td>
</tr>
<tr>
<td>Sometimes</td>
<td>20.6% (21)</td>
<td>59.8% (61)</td>
<td>19.6% (20)</td>
<td>100% (102)</td>
</tr>
<tr>
<td>Almost Never/ Never</td>
<td>55.0% (11)</td>
<td>35.0% (7)</td>
<td>10.0% (2)</td>
<td>100% (20)</td>
</tr>
</tbody>
</table>

χ² = 20.262. df= 4. p<.001

Table 5.7 is consistent with the findings on denunciation in Table 5.4. Judges who believe that rehabilitation can "almost never/never" be achieved through a conditional sentence were significantly more likely to consider the impact that the sanction will have on public opinion, than those who perceive rehabilitation to be accomplished "always and usually". Clearly the pattern of results found in Tables 5.2, 5.3, and 5.4 which focus on denunciation are similar to those found in Tables 5.5, 5.6, and 5.7 on rehabilitation. This finding has implications for understanding judges' perceived role of denunciation and conditional sentences.
Revising the Question: To what extent can a conditional sentence be as effective as imprisonment?

It is possible that the similar pattern of relationships found for rehabilitation and denunciation are for different reasons. For instance, judges' views about rehabilitation and its association with their perceptions about the public's understanding of the conditional sentence can be explained by another factor -- the ability to achieve rehabilitation in the community. This explanation was explored by looking at four questions in the survey of trial court judges that address issues of the adequacy of community resources and supervision relating to treatment programs, presumably in order to achieve rehabilitation effectively.\textsuperscript{22} It is likely that judges who believe that rehabilitation can be achieved through the conditional sentence "always and usually" will respond that there are adequate resources, and supervision available. However if judges' views about rehabilitation and denunciation are both significantly related to these factors in the same pattern, then one must conclude that judges are not answering the questions on achieving the purposes of sentencing in the manner in which they were intended. Indeed the analyses demonstrate that across all four questions, the same pattern of results was found for both denunciation and rehabilitation.\textsuperscript{23}

\textsuperscript{22} The following questions were included in the survey, see appendix to this Chapter. Q9: "If you are considering a conditional sentence, are you able to find out what community resources are available and which might be appropriate for the case before you?" Q10: "Do you believe that conditional sentence orders are being adequately supervised in your area?" Q11: "Would you be inclined to use conditional sentences more frequently if there were more community and supervisory resources?" Q13: "Is the number of available treatment and other programs in your area adequate to support the use of conditional sentences?"

\textsuperscript{23} Denunciation by community resources: $\chi^2 = 16.83$, df= 4, $p<.001$; denunciation by supervision: $\chi^2 = 54.62$, df= 4, $p<.0001$; denunciation by more resources: $\chi^2 = 9.57$, df= 2; $p<.01$; denunciation by programs: $\chi^2 = 27.43$, df= 4, $p<.001$. Rehabilitation by community resources: $\chi^2 = 21.644$, df= 4, $p<.001$; rehabilitation by supervision: $\chi^2 = 33.03$, df= 4, $p<.001$; rehabilitation by more resources: $\chi^2 = 14.02$, df= 2.
Judges certainly make distinctions between the purposes of denunciation and rehabilitation -- as shown in Table 5.1 and confirmed by the analysis of differences between means. However the results found in Tables 5.5, 5.6 and 5.7 also suggest that judges are answering the question on the purposes achieved by the conditional sentence as whether or not they "like" the sanction and believe it can generally accomplish the various goals of sentencing. If this is the case, then one would expect that all purposes are highly correlated. Indeed, Table 5.8 reveals that judges' views on the ability of the sanction to accomplish all five purposes are relatively highly correlated. This suggests, then, that those who "like" conditional sentences will be more likely to respond that this sanction can accomplish any goal, compared to those who do not "like" the sanction.

**Table 5.8**

Correlation Matrix of Judges' Perceptions of the Ability of the Conditional Sentence of Imprisonment to Achieve the Purposes of Sentencing

<table>
<thead>
<tr>
<th>Purposes of Sentencing</th>
<th>Proportionality</th>
<th>Denunciation</th>
<th>Deterrence</th>
<th>Rehabilitation</th>
<th>Victim Reparation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proportionality</td>
<td>1.00*</td>
<td>.6502</td>
<td>.5365</td>
<td>.5250</td>
<td>.3593</td>
</tr>
<tr>
<td>Denunciation</td>
<td>.6502**</td>
<td>1.00</td>
<td>.6499</td>
<td>.4812</td>
<td>.3392</td>
</tr>
<tr>
<td>Deterrence</td>
<td>.5365</td>
<td>.6499</td>
<td>1.00</td>
<td>.5349</td>
<td>.3941</td>
</tr>
<tr>
<td>Rehabilitation</td>
<td>.5250</td>
<td>.4812</td>
<td>.5349</td>
<td>1.00</td>
<td>.4247</td>
</tr>
<tr>
<td>Victim Reparation</td>
<td>.3593</td>
<td>.3392</td>
<td>.3941</td>
<td>.4247</td>
<td>1.00</td>
</tr>
</tbody>
</table>

*All n's are between 441 and 449.  
**All p's<.01

Thus it is possible that when relationships are explored between judges' views of accomplishing different purposes of sentencing by a conditional sentence and their other beliefs about this sanction, explanations can be determined by how judges generally rate

\[ p<.001; \text{rehabilitation by programs: } x^2 = 9.81, \text{ df}=4, p<.05. \]
the sanction. The original question was "Do you think that you are able to set conditions for a conditional sentence that are as effective as a normal sentence of imprisonment in accomplishing each of the following principles or purposes of sentencing?" Judges appear to be answering the question as "Do you think that a conditional sentence can be as effective (with the appropriate conditions) as imprisonment?" Therefore a new variable was created to represent a single rating by judges -- incorporating all the purposes of sentencing -- as "low", "medium", and "high" support for the question.24 Tables 5.9, 5.10, and 5.11 below present these relationships with the revised variable.

---

24 A new variable on support was created by recoding answers to each sentencing purpose as "low", "medium" and "high". Across all sentencing purposes, low ratings were added together, followed by all medium, and high ratings. This new variable represented low, medium or high support for the statement that the conditional sentence can be as effective as prison.
Table 5.9

Judges' Views About the Ability of a Conditional Sentence To be as Effective (With Conditions) as Imprisonment and Their Views on the Public's Understanding of the Sanction

<table>
<thead>
<tr>
<th>Conditional Sentence can be as Effective as Prison...</th>
<th>Most &amp; Some of the Public*</th>
<th>Few &amp; None of the public</th>
<th>Don't know</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low Support</td>
<td>9.1% (13)</td>
<td>53.1% (76)</td>
<td>37.8% (54)</td>
<td>100% (143)</td>
</tr>
<tr>
<td>Medium Support</td>
<td>18.5% (32)</td>
<td>66.5% (115)</td>
<td>15.0% (26)</td>
<td>100% (173)</td>
</tr>
<tr>
<td>High Support</td>
<td>24.8% (29)</td>
<td>64.1% (75)</td>
<td>11.1% (13)</td>
<td>100% (117)</td>
</tr>
</tbody>
</table>

*No judge responded that the public understands the nature of the conditional sentence of imprisonment "all of the time".

$\chi^2 = 38.919, \text{ df}= 4, p < .001$

Table 5.10

Judges' Views About the Ability of a Conditional Sentence To be as Effective (With Conditions) as Imprisonment and Their Views on Public Support of the Use of the Sanction

<table>
<thead>
<tr>
<th>Conditional Sentence can be as effective as prison</th>
<th>All, Most &amp; Some of the public who are aware</th>
<th>Few &amp; None of the public who are aware</th>
<th>Don't know</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low Support</td>
<td>26.2% (37)</td>
<td>58.2% (82)</td>
<td>15.6% (22)</td>
<td>100% (141)</td>
</tr>
<tr>
<td>Medium Support</td>
<td>60.1% (104)</td>
<td>24.9% (43)</td>
<td>15.0% (26)</td>
<td>100% (173)</td>
</tr>
<tr>
<td>High Support</td>
<td>80.2% (93)</td>
<td>10.3% (12)</td>
<td>9.5% (11)</td>
<td>100% (116)</td>
</tr>
</tbody>
</table>

$\chi^2 = 87.961, \text{ df}= 4, p < .0001$
Table 5.11
Judges' Views About the Ability of a Conditional Sentence To be as Effective (With Conditions) as Imprisonment and Their Consideration of the Impact of the Sanction on the Public

<table>
<thead>
<tr>
<th>Conditional Sentence can be as effective as prison</th>
<th>All of the Time</th>
<th>Most/Some of the Time</th>
<th>Unknown</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low Support</td>
<td>21.0% (29)</td>
<td>59.4% (82)</td>
<td>19.6% (27)</td>
<td>100% (138)</td>
</tr>
<tr>
<td>Medium Support</td>
<td>18.6% (32)</td>
<td>70.9% (122)</td>
<td>10.5% (18)</td>
<td>100% (172)</td>
</tr>
<tr>
<td>High Support</td>
<td>14.0% (16)</td>
<td>52.6% (60)</td>
<td>33.3% (38)</td>
<td>100% (114)</td>
</tr>
</tbody>
</table>

χ² = 23.974, df= 4, p<.001

The results from Tables 5.9, 5.10, and 5.11 reveal that across all three questions, judges who hold low support-- compared to high support -- for the statement that conditional sentences can be as effective as imprisonment are significantly less likely to believe that the public understands and supports the use of conditional sentences. Judges who have low support also are significantly more likely to consider the impact that a conditional sentence has on the public. More broadly, these findings confirm that judges are answering the questions on purposes largely in terms of whether or not the sanction is viewed as effective as imprisonment.

A further analysis supports this interpretation of the question. The extent to which judges have imposed conditional sentences (at the time the survey was carried out) also related significantly to judges' ratings of the effectiveness of the conditional sentence to
accomplish each of the five purposes. It was found that judges' use of the conditional sentence -- as 'none', 'low', and 'medium/high' -- is related to their ratings of the conditional sentence effectively accomplishing each purpose of sentencing. Those who impose the sanction more frequently are significantly more likely to see the conditional sentence accomplishing each goal as effectively as imprisonment, compared to low users. Thus the more judges “use” the sanction, the more they believe it can generally accomplish the various goals of sentencing. Table 5.12 below illustrates this finding. Each cell percentage represents the proportion of respondents who believe that each purpose of sentencing can be achieved as effectively as imprisonment (“always and usually”) “Medium/high” users were more likely than other users to believe that each purpose of sentencing could be achieved with conditions as effectively as imprisonment “always and usually”.

Table 5.12
Judges' Use of Conditional Sentences and Their Views That Each Purpose of Sentencing Can be Achieved (With Conditions) as Effectively as Prison “Always and Usually”

<table>
<thead>
<tr>
<th>Use of Conditional Sentence</th>
<th>Proportionality</th>
<th>Denunciation</th>
<th>Deterrence</th>
<th>Rehabilitation</th>
<th>Victim Reparation</th>
</tr>
</thead>
<tbody>
<tr>
<td>None</td>
<td>17.4%</td>
<td>21.7%</td>
<td>17.4%</td>
<td>47.8%</td>
<td>30.4%</td>
</tr>
<tr>
<td>Low (1-10)</td>
<td>43.4%</td>
<td>27.2%</td>
<td>31.1%</td>
<td>66.1%</td>
<td>54.7%</td>
</tr>
<tr>
<td>Medium/High (11+)</td>
<td>62.6%</td>
<td>45.1%</td>
<td>39.8%</td>
<td>80.1%</td>
<td>66.5%</td>
</tr>
</tbody>
</table>

*Proportion indicating that the purpose of sentencing can be achieved*

25 'Low' use of conditional sentences included 1-10 times, and “medium” and “high” were combined to include 11 times or more.
Court of Appeal Judges’ Views and the Denunciation-Rehabilitation Dichotomy

While findings from the survey of trial court judges have not resulted in a consistent pattern for the purpose of denunciation, interviews with Appeal Court judges provide some further evidence that judges’ views on denunciation play a significant role in their thinking about conditional sentences. One interview question -- with two subsections -- captures the essence of the relationship between judges’ views on denunciation and their perceptions about public understanding and support of conditional sentences, and the role of the public in judicial sentencing decisions.

All Court of Appeal judges were told that “[o]ne of the issues that I’m interested in is public attitudes towards different punishment....Do you think that members of the public will see a conditional sentence as being effective in denouncing crime? Or do you think that they will see imprisonment as more appropriate in accomplishing denunciation?” Secondly, “how do you think that the public will react towards this new punishment? Do you think that they will see it as a ‘let off’? Do you think that it depends upon the nature of the offence?” As Court of Appeal judges answered these questions, they also revealed their perceptions about public understanding and support of the conditional sentence, and its impact on public opinion.

All seven Court of Appeal judges responded that members of the public see and/or will see the conditional sentence as being ineffective in accomplishing denunciation compared to imprisonment, and that the public will see the sanction as a ‘let off’. Four out of seven judges specifically mentioned the nature of the offence as a factor affecting the public’s perception of a conditional sentence as being a ‘let off’. The two questions are related and were asked in sequence. Thus the responses of each question
will be merged and presented below. One judge (#1) remarked

No. They [the public] don’t have enough knowledge; most aren’t even aware what it is. Everyone understands prison. It gets the job done.... Yes [they see it as a ‘let off’]. Even the name is a bit of a fraud. It is misleading and a fraud to the public. They [Parliament] did not think about it too much. This is a problem. We should be more open about what we are trying to do. This is one of my concerns....

Judge #4 responded by emphasizing public understanding and support:

Yes. Prison will be seen as more denunciatory. I think that the public doesn’t think about these things with a lot of precision. Prison is punishment and everything else is letting off. With break and enter, they may see it [a conditional sentence] as appropriate but see it as always bad, and prison is the most appropriate if they had the choice.

Finally, another Court of Appeal judge (#5) reinforced these views:

The prison [will be seen by the public] as more severe and more denunciatory. I don’t mean to say that the conditional sentence is not effective, but not as effective [as imprisonment]. I’m skeptical of public acceptance, but as the public learns more, they see it as more acceptable, but not at the same level as imprisonment. They see it as a bit of both [a ‘let off’ and it depends upon the nature of the offence]. I’m sure that for the most minor offences they will see a conditional sentence as acceptable. But as you go up the scale of seriousness of the offence, they view it as a let off.

Thus the interviews with Court of Appeal judges presented above demonstrate that their views about denunciation are associated with their perceptions about the public’s understanding and acceptance of the sanction for various offences. Two Court of Appeal judges who were interviewed also consider public attitudes towards conditional sentences. For example, one judge stated (#7),
They [the public] would not see a conditional sentence as denunciatory. For example, in impaired driving causing bodily harm or death, they would see it as a slap on the wrist....They see it is as a let off. For first time or second or even a young person, there is not a problem with it, so long as there are adequate terms. The public don't like to pick up the tab though, like having to pay a $500 deductible.

In examining the data on trial court judges as a whole, it is apparent that those judges who have low support -- compared to high supporters -- for the view that the conditional sentence can be as effective as prison are less likely to believe that the public understands and supports the sanction. Not surprisingly, low supporters are more likely to consider the impact of the sanction on the public. Clearly the interview data and results found in Table 5.1 suggest that one reason for low support is their belief in the lack of denunciatory power of a conditional sentence. The findings also demonstrate that Court of Appeal judges who see the lack of denunciatory power of a conditional sentence has an impact, to some extent, on their views about public acceptability of the sanction.

**Conceptualizing the Conditional Sentence: Perceptions of the Public’s Understanding of the Difference Between a Conditional Sentence and Probation**

As mentioned earlier, an important issue is the extent to which the conditional sentence is viewed different from other sanctions within the community, such as probation. Cases such as *Brady* (Alta. C.A.) signify the concern by some courts that the conditional sentence is not different from probation (see also Roberts et. al 2000). Chapter 3 has also shown that members of the public do not make a distinction between the conditional sentence and probation. Since the sanction is to be understood as a replacement for imprisonment, it is supposed to be distinct from other ‘intermediate’ or
'community' sanctions. But if judges believe that it can be as effective as prison -- and thus equivalent and interchangeable -- then it is likely that they will have confidence in the public's views. In contrast, it is expected that low supporters of the belief that the conditional sentence can be as effective as prison will likely relate to their skepticism about the public's understanding of the sanction.

Therefore in the present survey of trial court judges, respondents were asked, "Do you think the general public can be made to understand the difference between a conditional sentence and a probation order?" and "Do you think that a conditional sentence has a different impact on an offender than a probation order with the same conditions?" These results are presented in Tables 5.13 and 5.14.
Table 5.13

Judges’ Views About the Ability of a Conditional Sentence to be as Effective (With Conditions) as Prison and Their Perceptions That the Public Can be Made to Understand the Difference Between the Conditional Sentence and Probation

<table>
<thead>
<tr>
<th>Conditional Sentence can be as effective as prison</th>
<th>All and Most of the Public</th>
<th>Some of the Public</th>
<th>Few and None of the Public</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low Support</td>
<td>7.2% (10)</td>
<td>29.7% (41)</td>
<td>63.0% (87)</td>
<td>100% (138)</td>
</tr>
<tr>
<td>Medium Support</td>
<td>36.3% (62)</td>
<td>31.0% (53)</td>
<td>32.7% (56)</td>
<td>100% (171)</td>
</tr>
<tr>
<td>High Support</td>
<td>63.2% (74)</td>
<td>23.1% (27)</td>
<td>13.7% (16)</td>
<td>100% (117)</td>
</tr>
</tbody>
</table>

χ² = 102.872, df= 4, p<.001

Table 5.14

Judges’ Views About the Ability of the Conditional Sentence to be as Effective (with Conditions) as Prison and their Perceptions That the Public Believe That a Conditional Sentence Has a Different Impact on Offenders Than a Probation Order With Same Conditions

<table>
<thead>
<tr>
<th>Conditional Sentence can be as effective as prison</th>
<th>Definitely and Probably “Yes”</th>
<th>Definitely and Probably “No”</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low Support</td>
<td>31.1% (41)</td>
<td>68.9% (91)</td>
<td>100% (132)</td>
</tr>
<tr>
<td>Medium Support</td>
<td>75.6% (121)</td>
<td>24.4% (39)</td>
<td>100% (160)</td>
</tr>
<tr>
<td>High Support</td>
<td>88.3% (98)</td>
<td>11.7% (13)</td>
<td>100% (111)</td>
</tr>
</tbody>
</table>

χ² = 100.562, df= 2, p<.001
As expected, low supporters -- judges who believe that a conditional sentence cannot be as effective as imprisonment -- are significantly less likely to believe that the public understands the difference between probation and the sanction than high supporters. Most supporters indicate that they believe that the public believes that a conditional sentence does not have a differential impact on an offender than probation, compared to high supporters.

Similarly, Court of Appeal judges were asked whether they think “the public sees a difference between a conditional sentence and probation”. All seven judges who were interviewed responded that they did not think members of the public see the difference -- both conceptually and practically. For example, Judge #3 stated

I don’t know, I don’t think that they [the public] understand a conditional sentence. It is even an odd word. I have no sense of how they feel. Some lawyers don’t know [the difference between a conditional sentence and probation] and a lot of Crowns don’t see a difference, as presently given.

Court of Appeal judge #6 reinforced the same sentiment: “They [the public] don’t. I think that they don’t see a difference. In fact, there is even confusion among lawyers.” Another responded stated that “No, they [the public] don’t see the difference. I think that the public lumps them into one category -- letting off.” Finally, another judge (#1) articulated the role of Courts of Appeal in ensuring the distinction between a conditional sentence and probation:

No, they [the public] don’t understand it. We’ve been lucky, we haven’t seen any public outrage yet. Judges may be doing a good job here. If a conditional sentence is imposed in inappropriate cases and the Court of Appeal does nothing, then there will be a problem. The Court of Appeal must make sure of this [the appropriate use of conditional sentences by trial court judges].
Clearly when examining trial court judges' responses, low support for the view that a conditional sentence can be as effective as imprisonment is significantly associated with their beliefs about the difference between probation and a conditional sentence. But accomplishing the purpose of denunciation likely plays some role in judges' beliefs about the lack of difference between probation and conditional sentences. To some extent, judges' perceived lack of equivalence between achieving denunciation through the conditional sentence compared to prison explains why some judges are less likely to believe that the public understands the difference between probation and conditional sentences. It is likely that the sanction is thought of less as "imprisonment" and more like a community sanction.

In addition, attaching the conditions to a conditional sentence that are similar to terms of probation (Roberts et al. 2000) reinforces the propensity to see the conditional sentence as a community sanction like probation. Achieving denunciation through the prison, and rehabilitation in the community, make up some part of judges' views about conditional sentences compared to probation. Thus at a conceptual level, it is apparent that the conditional sentence is not perceived to be seen by the public as equivalent, and therefore universally interchangeable with prison sentences.

**Conceptualizing the Conditional Sentence: Perceptions of Restricting An Offender's Liberty After Breach of a Conditional Sentence**

Some questions were included in the survey about judges' views about breaches of conditional sentences of imprisonment. While the survey was conducted before legal
changes were made to the administration of breach of a conditional sentence (section 742.6(1)(a)), respondents were asked about the appropriateness of denying bail, and sending the offender automatically to jail for a breach. More specifically, the questions posed to judges included: "Do you think an offender who breaches a conditional sentence should be denied bail?" and "Do you think an offender who breaches a conditional sentence should be automatically sent to prison to serve the balance of the sentence?"

Tables 5.15 and 5.16 present the results of the relationship between judges’ views about the ability of the conditional sentence to be as effective as imprisonment and their attitudes towards denial of bail and automatic prison for breach of a conditional sentence.

Table 5.15

Judges’ Views About the Ability of a Conditional Sentence to be as Effective (With Conditions) as Prison and Their Views on Denying Bail for Breaching

<table>
<thead>
<tr>
<th>Conditional Sentence can be as effective as prison</th>
<th>Yes, in All cases</th>
<th>Yes, in Most cases</th>
<th>Yes, in Some, and a Few cases &amp; No</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low Support</td>
<td>23.3% (31)</td>
<td>48.9% (65)</td>
<td>27.8% (37)</td>
<td>100%  (133)</td>
</tr>
<tr>
<td>Medium Support</td>
<td>14.2% (24)</td>
<td>45.0% (76)</td>
<td>40.8% (69)</td>
<td>100%  (169)</td>
</tr>
<tr>
<td>High Support</td>
<td>10.4% (12)</td>
<td>45.2% (52)</td>
<td>44.3% (51)</td>
<td>100%  (115)</td>
</tr>
</tbody>
</table>

$\chi^2 = 12.50$, df=4, $p<.05$
Table 5.16
Judges’ Views About the Ability of a Conditional Sentence to be as Effective (With Conditions) as Prison and Their Views on Automatic Prison for Breaches

<table>
<thead>
<tr>
<th>Conditional Sentence can be as effective as prison</th>
<th>Yes, in All cases</th>
<th>Yes, in Most cases</th>
<th>Yes, in Some, and a Few cases</th>
<th>No</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low Support</td>
<td>11.0% (17)</td>
<td>39.0% (60)</td>
<td>29.2% (45)</td>
<td>20.8% (32)</td>
<td>100% (154)</td>
</tr>
<tr>
<td>Medium Support</td>
<td>12.7% (18)</td>
<td>45.1% (64)</td>
<td>28.2% (40)</td>
<td>14.1% (8)</td>
<td>100% (142)</td>
</tr>
<tr>
<td>High Support</td>
<td>26.5% (36)</td>
<td>45.6% (62)</td>
<td>20.6% (28)</td>
<td>7.4% (10)</td>
<td>100% (136)</td>
</tr>
</tbody>
</table>

χ² = 33.861, df= 6, p<.001

Low supporters were significantly more likely to prefer denial of bail compared to high supporters. It is not surprising, therefore, that low supporters are less likely to prefer that breachers be sent back to jail automatically -- it results in the same consequence as automatic jail. One might argue that judges’ preferences for prison after a breach has occurred may be for the purpose of incapacitation. However, it is unlikely that an offender who requires separation from society from the outset would be sentenced by a judge to a conditional sentence because of the provision relating to “danger to the community” (section 742.1(b)).

Appeal Court judges were asked, “Do you think that conditions such as house arrest and electronic monitoring would make conditional sentences more acceptable to you in cases of serious violence or sexual offences? Why or why not? Does it relate to a particular goal of sentencing like incapacitation? Denunciation? Deterrence?
Rehabilitation? Victim reparation?” Immediately following this question, they were asked “If incapacitation is a consideration in a case, do you think that a conditional sentence is an appropriate punishment?”

Four out of seven judges (Judges #1, 3, 5, and 7) replied that restrictive conditions such as house arrest and electronic monitoring are appropriate conditions to be attached to conditional sentences in order to accomplish denunciation more effectively. The other three judges were fundamentally opposed to these restrictive conditions, preferring to provide offenders with additional treatment through probation (Judge #2) or a conditional sentence. Finally, all seven Court of Appeal judges responded that if incapacitation is a consideration in a case, then prison is required rather a conditional sentence of imprisonment. Some examples are provided below.

Judge #3 stated that

Electronic monitoring, I don’t know about...House arrest. yes, to me. and probably to the community [would be more acceptable in cases of serious violence or sexual offences]. Denunciation is the issue.... It [the conditional sentence] has some teeth. and the “teeth” is not incapacitation but denunciation -- it makes it more acceptable to the public....

Following this response, the same judge continued to talk about the goal of incapacitation and the conditional sentence, while also providing insight on appellate decision-making:

As a practical point, we try to reach a consensus. If dangerous, then should be imprisoned....House arrest might be a tough term to get a third judge to agree with the other two. Here is something about appellate decision-making. It is give and take. you need to compromise. I might not agree with every term of the sentence. but it is give and take, so I sign off. I have to really disagree and feel strongly about it to dissent. Now house arrest is one way of getting a dissenter to agree [to a conditional sentence].

The quote above most likely refers to achieving denunciation -- by a conditional sentence
or imprisonment -- and the condition of house arrest is one means by which judges can agree to a conditional sentence. Clearly the importance of the role of denunciation and other expressive purposes are reinforced when Court of Appeal judges are thinking about restrictive conditions. Judge #1 explains

[Conditions such as house arrest and electronic monitoring] gives us another alternative, and we can demonstrate that there was a significant element of denunciation and general deterrence....The conditions relate to the three principles of general deterrence, denunciation, and reinforcing accountability in offenders, not incapacitation. The conditions attached relate to general deterrence by showing there are real consequences, and denunciation by showing the community as a whole that values are reinforced. They also reinforce accountability in offenders....

It is apparent that Court of Appeal judges’ views about accomplishing denunciation through a conditional sentence of imprisonment are related to their other attitudes about the need for restrictive conditions. House arrest and electronic monitoring provide Court of Appeal judges with ways to achieve denunciation, which to some extent, meet the levels of restrictiveness – in both instrumental and expressive ways -- found through the use of the imprisonment. While the interview data do not directly address the issue of an offender who breaches a conditional sentence, they do serve as illustrations of judicial views about the lack of equivalence in achieving denunciation with the sanction compared to imprisonment.

On the whole, the findings demonstrate that those judges who do not believe that a conditional sentence is as effective as imprisonment prefer that offenders who breach should be denied bail and therefore brought to jail. This suggests that judges do not view the conditional sentence and imprisonment as equally restrictive. The interviews imply that one reason that judges desire further restrictions of liberty through jail is to send a
denunciatory message about following required conditions.

*Interchangeability Across All Offences?: Denunciation and Judges’ Views About the Suitability and Unsuitability of Offences for a Conditional Sentence*

The last relationship explored is one of the most controversial issues surrounding conditional sentences -- the imposition of conditional sentences for violent offences. It was clear from the last Chapter that denunciation is one of the main purposes for sentencing violent offenders to prison. But to what extent do judges see the conditional sentence as an appropriate response for violent offences? The present analysis on the suitability and unsuitability of offences is based upon a comparison of judges who believe that denunciation can be achieved by a conditional sentence “always/usually”, “sometimes” or “almost never/never”.26

Trial court judges were told, “Parliament has laid down certain criteria to be fulfilled before a conditional sentence may be imposed. For example, offenders must have been sentenced to a term of imprisonment of less than two years. Having taken these criteria into account, is there any kind of offence that you believe is particularly suited to a conditional sentence?” “Is there any kind of offence that you believe is particularly unsuited to a conditional sentence?” Judges were free to mention any offence, including both specific and whole categories. If a conditional sentence and imprisonment are viewed by judges as equivalent and interchangeable, then one would expect that they will not mention violence as being unsuitable. However it is possible, on the other hand, that judges might mention violence as both suitable and unsuitable, depending upon the nature

---

26 The variable on judges’ views about denunciation is the original one. Recall that judges’ responses to all sentencing purposes (that each purpose can be achieved by a conditional sentence “always/usually”).
or type of violence. For instance, a respondent might mention that "violence" is unsuitable for a conditional sentence and assault is suitable. Violent and property offences are not mutually exclusive. In general, violence was dramatically more likely to be seen as unsuitable than property for all groups.

Table 5.17

Judges’ Views About Denunciation and Their Attitudes Towards the Unsuitability of Violent Offences for Conditional Sentences

<table>
<thead>
<tr>
<th>Denunciation can be achieved by a conditional sentence...</th>
<th>Violence mentioned as unsuitable</th>
<th>Violence not mentioned as unsuitable</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Always/Usually</td>
<td>59.5% (94)</td>
<td>40.5% (64)</td>
<td>100% (158)</td>
</tr>
<tr>
<td>Sometimes</td>
<td>70.3% (104)</td>
<td>29.7% (44)</td>
<td>100% (148)</td>
</tr>
<tr>
<td>Almost never/Never</td>
<td>72.5% (103)</td>
<td>27.5% (39)</td>
<td>100% (142)</td>
</tr>
</tbody>
</table>

χ² = 6.722. df = 2, p < .035

"sometimes", and "almost never") were subsequently pooled to create a single rating. See note 24.
Table 5.18
Judges’ Views About Denunciation and Their Attitudes Towards the Unsuitability of Property Offences for Conditional Sentences

<table>
<thead>
<tr>
<th>Denunciation can be achieved by a conditional sentence...</th>
<th>Property Offences mentioned as unsuitable</th>
<th>Property Offences not mentioned as unsuitable</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Always/ Usually</td>
<td>3.2%</td>
<td>96.8%</td>
<td>100%</td>
</tr>
<tr>
<td>(5)</td>
<td>(153)</td>
<td>(158)</td>
<td></td>
</tr>
<tr>
<td>Sometimes</td>
<td>6.8%</td>
<td>93.2%</td>
<td>100%</td>
</tr>
<tr>
<td>(10)</td>
<td>(138)</td>
<td>(148)</td>
<td></td>
</tr>
<tr>
<td>Almost never/ Never</td>
<td>9.2%</td>
<td>90.8%</td>
<td>100%</td>
</tr>
<tr>
<td>(13)</td>
<td>(129)</td>
<td>(142)</td>
<td></td>
</tr>
</tbody>
</table>

$\chi^2 = 4.78, \text{ df} = 2, p < .01$

Tables 5.17 and 5.18 reinforce the hypothesis that judicial attitudes about the difficulty of achieving denunciation with a conditional sentence is related to their views about the appropriateness of property and violent offences for a conditional sentence. As expected, those judges who were less confident in the ability of a conditional sentence to achieve denunciation were more likely to mention property offences as unsuitable, compared to those who see the conditional sentence as “always” and “usually” denunciatory. Unsuitable offences were described as “violent”, “sexual assault”, “family violence”, “where injury occurs”, “child sexual assault”, “hurt women/children”, “weapons offences”, “threat of violence”, and “robbery”.

Looking at judges’ views about the suitability of offences, it was expected that those who believe that the conditional sentence lacks denunciatory power will be less likely to respond that violent offences are suitable and more likely to say that property
offences are suitable. When the same relationship was explored, however, significant results were not found. Some suitable offences -- of a violent nature -- were mentioned including “historic sexual assault”, “minor violence”, “minor sexual assault”, “minor spousal assault”, and “robbery”. Clearly judges make distinctions among individual offences, even within the “violent” category.

In order to confirm whether these findings are related to denunciation and not to other purposes, the relationship between judges’ views about rehabilitation and the unsuitability and suitability of violent and property offences were explored. Non-significant results were found across all four questions. Thus the data demonstrate that denunciation is an important dimension in their views about the kinds of offences that are inappropriate for conditional sentences.

**Court of Appeal Judges’ Views about the Suitability and Unsuitability of Offences**

Two questions on denunciation and the suitability of offences were asked of Appeal Court judges. First, they were asked, “When you see denunciation as an important sentencing goal in a case, do you think that a conditional sentence is appropriate, and why? If no, why? In which cases might it not be?” Following this question and intimately related, they were asked, “When you see denunciation as an important sentencing goal in a case, do you think that imprisonment is appropriate, and why? If no, why? In which cases might it not be?”

---

27 **Denunciation by unsuitability of violence:** $\chi^2 = 0.63$, df= 2, not significant; Denunciation: $\chi^2 = 1.59$, df=2, not significant.

28 **Rehabilitation by unsuitability of violence:** $\chi^2 = 4.404$, df= 2, not significant; Rehabilitation by unsuitability of violence: $\chi^2 = 1.148$, df= 2, not significant; Rehabilitation by suitability of violence: $\chi^2 = 0.305$, df= 2, not significant. Rehabilitation by suitability of property: $\chi^2 = 3.673$, df= 2, not significant.
All seven judges who were interviewed stated that when denunciation is important in a case, a conditional sentence can be appropriate, depending upon the nature of the offence. In particular, the extent to which there is violence will dictate whether a conditional sentence or imprisonment is appropriate. Six out of the seven judges stated that they do not exclude whole categories of offences from eligibility of a conditional sentence. However they conceptualize the relationship between denunciation and violence as interacting at various levels. For instance, Judge #1 explained:

Let's take sexual assault of children for example. The courts have said that denunciation is important in these cases, and denunciation is important in all sexual assaults, but more so if it is a case of breach of trust. In a recent case of sexual assault (not as severe as rape) in a breach of trust, denunciation was important. The defence argued for house arrest, and I said that a conditional sentence was not appropriate. There are different degrees of sexual assault. Denunciation is important and a conditional sentence is a hard sell.

Judge #2 stated,

...I don’t exclude whole categories. I do acknowledge that there are different levels of violence and sexual assault. Minor sexual assault, i.e. touching, it is appropriate. But sexual assault with breach of trust, I disagree, likely not. So I look at whether violence causes serious harm vs. less serious harm.

Judge #7 reinforced notions of different degrees of violence and denunciation,

For example, in a serious sexual assault, a conditional sentence is not appropriate. It does not emphasize denunciation. A robbery with violence, it is difficult to persuade people that a conditional sentence is warranted. In a regular robbery even, there is that threat of violence. I personally would think that the public would disagree unless there was a lot of monitoring and reeducating of the offender....

**Equivalency and the Nature of the Offence**

The next set of questions asked Court of Appeal judges “Do you think that a
conditional sentence of imprisonment is equivalent, or can it be made equivalent to a sentence of imprisonment? Why or why not? Does it depend upon the nature of the offence?" Six out of the seven judges stated that they do not believe that the conditional sentence is or can be made equivalent to a sentence of imprisonment, as well as that the nature of the offence is an important consideration. Judge #2 articulated the relationship between denunciation and the nature of the sanction in talking about the appropriateness of a conditional sentence or imprisonment:

No, unequivocally [the two are not equivalent]. There is no loss of liberty in a conditional sentence. The only difference with probation is that breach immediately kicks in. You must be in jail to know the experience. It is not like being in your home. No guards, no inmates in your house to eat and live with. It is qualitatively different. Yes, it depends upon the nature of the offence and the offender whether jail is appropriate....If the crime is serious, or if risk or need for general deterrence, then jail. If crime involves money, then I try to give a break depending upon the circumstances.

Judge #7 reinforces the lack of equivalence of a conditional sentence and imprisonment:

For my part, a conditional sentence is not equivalent to prison, or from the standpoint of the public, and that's what counts. The credibility of the system rests on public perception. If more serious, then they are not equivalent. For a first time offence, or even second, a conditional sentence is similar to a suspended sentence. The person deserves a break; but not prison to begin with.

Thus far the two quotes illustrate that the sanctions are perceived by judges to lack equivalence, to be qualitatively different, and in fact closer to other community sanctions like probation than prison. In conjunction with other quotes presented in this Chapter, it is clear that a fundamental factor for Court of Appeal judges in deciding on the conditional sentence depends upon the nature of the offence. Implicit in their responses is that the
'nature' of the offence is made up of qualities of differing levels of seriousness and violence. This corresponds to the need to accomplish an expressive goal like denunciation or general deterrence. Thus the conditional sentence is not assumed to be appropriate across all purposes and offences.

**The Multidimensionality of Punishment and the Role of the Offence**

The final set of questions reinforces the data presented in this Chapter that judges see sanctions as multidimensional. Court of Appeal judges were asked whether other sanctions such as community service orders and fines could be made equivalent and substituted for some term of imprisonment. Judges were told, "The Criminal Code now suggests that judges should consider non-prison sanctions as much as possible and when appropriate. One way of using prison less is to try to make non-prison sanctions somewhat equivalent. Do you think that a community service order/fine could be made equivalent and substituted for some term of imprisonment? Why or why not? Does the offence matter?"

Like the conditional sentence, all seven judges stated that the extent to which community service orders and fines are interchangeable with sentences of imprisonment depends, to some extent, upon the nature of the offence. However, all seven judges responded that in some instances, the community service is more easily substitutable for a sentence of imprisonment than the fine. These findings are consistent with the results in Chapter 2 relating to public sensibilities about their preference towards substituting a community service order, compared to fines, for sentences of imprisonment. Judge #4 articulated the qualities of different sanctions, and the multiple dimensions of punishment
-- the purpose, sanction and the nature of the offence.

It [the community service order] can be substituted for, but I don’t see it as equivalent. You can punish effectively without sending them to jail always, yes. Sexual assault, no. It is a case of degrees, and it depends upon the offence and the circumstances. A fine gets tricky. The means of the offender is important. When it is a regulatory offence, then yes, a fine is okay. But criminal offences, fines are less appropriate. It is different with fines and CSOs (community service orders).

This response is a clear illustration that judges conceptualize punishment as multidimensional. It reinforces judicial perceptions about the connections and intersections they see among the sanction, nature of the offence, and sentencing purpose. Clearly some sexual assaults are viewed as being more serious than others, and thus require different purposes such as denunciation. These Courts of Appeal judges make distinctions between the qualities of a community service order and a fine within a broader context of the prison. the qualities of the offence, and presumably the underlying purpose of sentencing that is seen as required for different offenders.

The next response presented is by Judge #2 who answered the questions on interchangeability of community service orders and fines for sentences of imprisonment. However, the issue about interchangeability was viewed and discussed within the same context as conditional sentences. In this way, it appears that Judge #2 sees the conditional sentence as a ‘community sanction’ like a community service order or a fine -- not as a ‘sentence of imprisonment’. Again, the sanctions are viewed as qualitatively different from each other, and all different from prison -- the dichotomy of the ‘community’ and ‘the prison’. The Court of Appeal judge states that

I am open to any form of an exaggerated form of probation, like a community service order attached to probation. It is far more honest [than a conditional sentence]. A conditional sentence is a political judgment in light of victims needs
and the expense of imprisonment. A conditional sentence is a political tool to divert people from reality. A CSO, it depends upon the offender and the service. I think that it should be used more, but we need more support services and probation officers. It is an effective alternative to prison in some cases. But I would need Parliament to give me some guidelines, and judges might do it more. I'm nervous about the fine because of default problems. I would really make sure that people can afford it. We don't spend enough time thinking about its impact on offenders. The fine is not right for all offences, especially an offence dealing with money. I could afford to pay more than someone else. At the same time, a middle manager paying a big fine might really hurt. But a fine and a CSO are different, both in terms of severity and qualitatively. A CSO is based upon you being punished in terms of time; with a fine, time won't make a difference in trying to find money to pay.

Finally, Court of Appeal judges were told that “some have argued that a range of community punishments are easily interchangeable with imprisonment as long as they are severe enough. Would you agree that punishments are easily interchangeable? Why or why not?” This set of responses confirm that Court of Appeal judges do not see sanctions as easily interchangeable, and do not assume that severity of the sanction is the focal point of sentencing as one might expect. Rather punishment is viewed as multidimensional.

Judge #5, for instance, responded in simple terms: “No, prison is not equivalent to these sanctions. You can't compare.” On the other hand Judge #3 appears to be the only respondent to favour interchangeability by focusing exclusively on the severity of the sanction. However it is interesting that his/her comments incorporate the quality of denunciation is explaining different sanctions. The respondent stated:

I'm not sure that interchangeability is the right word. But sanctions would be seen to be more interchangeable if they were more severe and onerous. For example, CSOs. If heavy, it does send a message, or a conditional sentence coupled with CSOs. It sends the right message.
Finally Judge #1 suggests a shift away from prison as the central focal point in sentencing:

I would be willing to be persuaded through data [that interchangeability of community punishments with prison is possible so long as severe enough] if it accomplishes the same goals, but you may have a tough time proving that. There is so little data on the effectiveness of prison, we know that first offenders are likely to reoffend if you put them in jail. We know that it is bad. The public is conditioned that the only meaningful punishment is prison. We don’t explore anything else. Scandinavian countries are good at this. Community punishments are appropriate and prison is inappropriate.

In this way, Court of Appeal judges do not view interchangeability and ‘intermediate’ sanctions as appropriate across different sanctions, offences, and purposes -- sanctions are assumed to be multidimensional. Community service orders and fines were seen, for the most part, as lacking equivalence to imprisonment and qualitatively different from each other.29

Conclusion

An exploration of judicial sensibilities about conditional sentences and imprisonment is one way to understand issues about interchangeability, the multiple dimensions of sanctions, and the connections between prison and the expressive functions of punishment. The present analysis began with the hypothesis that the purpose of denunciation plays an important role in judges’ sensibilities about the conditional sentence compared to imprisonment. It was argued that judges see the conditional

29 Another question asked Court of Appeal judges “As I have mentioned, some have argued that a range of community punishments are easily interchangeable with imprisonment as long as they are severe enough. Would you agree that punishments are easily interchangeable? Why or why not?” (Q12). Five out of seven Court of Appeal judges responded that sanctions are not easily interchangeable with imprisonment because it depends upon the nature of the offence -- violent and sexual offences in particular. Some judges also
sentence as lacking denunciatory power when compared to prison. It was proposed that their perceptions about denunciation would be related to some other views about conditional sentences. Other attitudes include the public’s lack of understanding and support of the sanction, as well whether individual judges themselves consider the impact that a conditional sentence has on the public. The issue of denunciation was also hypothesized to relate to judges’ conceptualizations of the conditional sentence -- the sanction is seen as similar to probation and lacking equivalence to prison. Finally, it was argued that their views about the lack of denunciatory power of denunciation would be related to their attitudes towards the inappropriateness of violence for conditional sentences.

The findings presented in this Chapter point to the importance of both trial and appeal court views about denunciation and its explanatory power for understanding other attitudes they hold about interchangeability and the equivalence of the conditional sentence and prison. However when exploring trial court judges views about the public’s understanding and support of conditional sentences, and judges’ perceptions about the difference between a conditional sentence and probation and prison, it was difficult to make conclusions about the role of denunciation per se. As it was demonstrated earlier in Tables 5.2 through 5.7, trial court judges had a tendency to answer the question about purposes as whether the conditional sentence was as effective as prison.

Nevertheless trial and Court of Appeal judges saw a lack of denunciatory power in the conditional sentence of imprisonment (Table 5.1). Court of Appeal judges in particular, responded to all questions by relating their views about denunciation to their stated that it depends upon the nature or class of the offender.
beliefs that the public do not understand and support the use of conditional sentences. do not see a conditional sentence as similar to probation or equivalent to imprisonment, and finally that some violent offences are seen as unsuitable by judges for conditional sentences. Thus it can be concluded that on the whole, denunciation is a substantial -- but not universal -- explanation for understanding judges’ sensibilities about the lack of equivalence between a conditional sentence of imprisonment and prison, as well as their experiences with various challenges surrounding the concept of interchangeability.

It is apparent that understanding punishment as multidimensional -- rather than a simple belief that sanctions are only different on a single dimension of severity -- provides an effective context for examining judicial sensibilities towards conditional sentences. First it is clear that judges saw the conditional sentence as significantly more capable to meet rehabilitative ends than to denounce, when compared to prison (Table 5.1). Clearly this finding stands in contrast to Morris and Tonry’s (1990) conceptualization of interchangeability. The authors do not address that there might be limits to interchangeability in terms of the purposes and functions that sanctions are seen as serving. The present findings reveal that judges do not view the conditional sentence as equivalent and easily interchangeable with a sentence of imprisonment across all purposes.

As mentioned throughout this thesis, there are a number of studies and

---

30 It is important to note that while Morris and Tonry (1990) initially do not address the purposes of sentencing within their proposals for interchangeable sanctions, Tonry’s later work (1998) addresses the functions of punishment within an interchangeability model. Tonry argues that “Instead, the focus would be on interchangeability of function. One penalty would be substituted for another if they both serve the same or alternately appropriate penological purposes in the circumstances” (original emphasis) (Tonry 1998: 292-3). However, he does not discuss the perceived connections among purposes, functions, and qualities of the offence and offender. See Tonry, M. “Interchangeability, Desert Limits and Equivalence of Function.” in von Hirsch, A. and A. Ashworth. 1998. Principled Sentencing: Readings on Theory & Policy. 2nd ed.
commentaries that have highlighted the severity of sanctions in order to create equivalencies and therefore substitute one penalty for a sentence of imprisonment. Writings on proportionality and desert in sentencing also underline the importance of severity of the sanction in order to achieve ‘fairness’ among offenders (von Hirsch 1976, 1998). But the findings in this Chapter demonstrate that judges see the conditional sentence as multidimensional -- having different qualities in addition to quantity.

To some extent, judges’ views about the difficulty of achieving denunciation through a conditional sentence and its relationship to their perceptions about the lack of public understanding and support of this sanction suggests that judges see another dimension to ‘punishment’ -- public sensibilities. While Chapter 2 examined public sensibilities about interchangeable sanctions, the present analysis of judges reveals the public’s views about what is acceptable, or penal culture, is considered by judges when they think about conditional sentences. As mentioned earlier in this Chapter, a variety of cases mention the importance of the public’s perceptions and acceptance of the imposition of punishments (Brady, Parker, W.(L.F.)). In addition, Garland (1990) underlines the salient role of the public in sentencing, compared to the operation of penal institutions:

By and large the institutions of punishment escape close scrutiny by the public. When new institutions are developed, or major reforms undertaken, it is politically requisite that these be represented in such a way as to gain the tacit support of the public -- or at least they must avoid incurring the opposition of those parts of the public which are politically active. However, once the institutional structure is established its repertoire of sanctions becomes the conventional one and it quickly becomes a routinized, technical matter. So long as

Oxford: Hart.
31See note 1.
the existing sanctions appear to convey a punitive effect in a manner which is broadly in keeping with current sensibilities, there tends to be limited moral interest in the details of how punishments are actually carried out. The whole focus of public interest and emotional intensity falls upon the court ritual and the declaration of sentence (Garland 1990: 73).

The findings also revealed that judges do not view a conditional sentence and imprisonment as equivalent on another two levels: when measured against probation, and when measured against the restrictive qualities of the prison. The denunciatory aspect of the prison -- mentioned along with the need for more restrictive conditions on the offender -- stands out as the distinguishing factor between the conditional sentence and probation on the one hand, and the prison on the other hand. The Court of Appeal of Alberta in Brady provided the first statement by the courts that the conditional sentence was similar to probation.32 Roberts et al. (2000) recently found in their study that public knowledge of the conditions attached to a conditional sentence had an impact on the level of public acceptability of the sanction. Furthermore, an analysis of the number and nature of conditions attached to probation orders compared to conditional sentences were similar (2000: 122). Indeed Manson (1997a) argues that whether the conditional sentence replaces probation or imprisonment will be affected by the difficulty of imposing conditional sentences when denunciatory statements of the courts are desired (1997a: 252).

More recent attention has been given to the nature of the conditions attached to conditional sentences in order to improve the ability to achieve a level of denunciation that is perceived as equivalent to prison. Therefore it is here that the quantitative and

qualitative dimensions of sanctions appear to interact. The case of Wismayer was one of the first pronouncements by the Ontario Court of Appeal in stating that "the objective of denunciation can be fully satisfied including strict conditions." More recently cases such as Smith (1999) reinforce the sentiments of the courts to distinguish the conditional sentence from probation, and make it equivalent to prison by imposing restrictive conditions such as curfews, house arrest, and electronic monitoring. However the defendant Smith served the sentence in custody in order to "adequately reflect the denunciation which must be accorded to this behaviour within our community" (Smith at 3, [1999] O.J. No. 2694).4

The final statement about denunciation, conditional sentences, and restrictive conditions has been made by the Supreme Court of Canada in responding to five conditional sentence appeals. In the lead case of Proulx, Chief Justice Lamer discussed the ability to achieve denunciation and general deterrence through the imposition of conditions such as house arrest.5 Two drug trafficking cases from the Saskatchewan Court of Appeal -- post-Proulx -- have carried through the imposition of restrictive conditions such as house arrest as a condition of the conditional sentence in order to achieve denunciation and general deterrence.6 Clearly denunciation underpins decisions by courts to impose restrictive conditions with a conditional sentence. Restrictive

---

34 See also R. v. Gagnon. (1998). 130 C.C.C. (3d) 194, when the Quebec Court of Appeal argued "... that a conditional one-year sentence of imprisonment adequately denounces appellant's unlawful conduct...if she is thereby deprived of her liberty in a significant manner for a sufficient period of time" at page 477. On the other hand, provinces such as Saskatchewan have been imposing conditions such as house arrest and electronic monitoring relatively frequently when compared to other provinces. For an analysis of case law relating to conditions attached to conditional sentences, see Renaud, G. "Conditions? Conditions! Conditions..." Paper presented to the Continuing Education Conference of the Ontario Court of Justice. June 17, 1999. London, Ontario.
conditions attached to a conditional sentence not only represent a dimension of the severity of punishment but also a qualitative aspect about the expression of condemnation of the behaviour.

Finally this Chapter demonstrated that judicial sensibilities toward achieving denunciation through a conditional sentence are based on the nature of the offence. The purpose of the sanction and the nature of offence appear to interact in these findings. The data showed that judges' views about denunciation -- at both trial and appeal court levels -- were significantly related to their attitudes toward the unsuitability of violent and sexual offences. Numerous cases relating to the ability of achieving denunciation for violent and sexual offences through a conditional sentence were highlighted much earlier in this Chapter. It appears that while the Supreme Court of Canada did not restrict any particular offences or categories from the eligibility of conditional sentences, the Court suggests restrictions based upon the ability of achieving denunciation and general deterrence for some offences where incarceration is the only way to express society's condemnation or deter in the future.37

Thus at all levels, judges see the complexities involved in substituting sanctions for sentences of imprisonment, rather than a simplistic view of creating equivalencies by focusing on the severity of sanctions. They see the sanction as multidimensional, having different qualities compared to imprisonment -- across different offences and purposes of sentencing. Not only does this examination tell us about judges' views about denunciation, the conditional sentence, and prison, but it provides a tool for

---

understanding their broader values and perceptions about what sanctions entail -- both about quantity and quality.
Chapter 5: Appendix A

Interview with Court of Appeal Judges on Conditional Sentences

1. What do you look at when receiving a conditional sentencing appeal?
   a) for instance, is the nature of the offence an important consideration? if it is property?
      if it is violent? sexual? If Yes, what is relevant about the offence?
   b) How do you assess dangerousness of the offender to the community?
   c) What do you think is accomplished through the development of a conditional sentence in Canada? What do you think is its purpose?
   d) What do you think is accomplished through a conditional sentence in terms of the goals of sentencing (denunciation; general deterrence; individual deterrence; assist rehabilitation; victim reparation; accountability)?

2. When you see denunciation as an important sentencing goal in a case, do you think that a conditional sentence is appropriate, and why? If no, why? In which cases might it not be?

3. When you see denunciation as an important sentencing goal in a case, do you think that imprisonment is appropriate, and why? If no, why? In which cases might it not be?

4. Are there any offender characteristics relevant in making your decision about a conditional sentence of imprisonment? Gender? Age? Past convictions?

5. Do you think that imprisonment has traditionally been seen by judges as the way to accomplish denunciation or general deterrence? -- the expressive goals of sentencing? Why or why not?

6. A number of Court of Appeal cases across the country have attempted to link
conditional sentences with accomplishing denunciation and general deterrence. Do you think that a conditional sentence of imprisonment can accomplish denunciation and general deterrence as effectively as imprisonment? Why or why not?

7. Do you think that there is much disagreement among the judiciary about this link between conditional sentences and denunciation and general deterrence? Why or why not?

8. There seems to be a lot of discretion for judges when it comes to the conditions attached to conditional sentences.

a) Do you think that conditions such as house arrest and electronic monitoring would make conditional sentences more acceptable to you in cases of serious violence or sexual offences? Why or why not? Does it relate to a particular goal of sentencing like incapacitation? denunciation? deterrence? rehabilitation? victim reparation?

b) If incapacitation is a consideration in a case, do you think that a conditional sentence is an appropriate punishment?

9. What do you think it is about conditional sentences that makes them different than other community punishments, like probation or community service orders? (Is it an issue of severity -- is it that a conditional sentence is more severe than let’s say probation; or is it that they different in terms of quality -- that they accomplish different goals etc.)?

10. Section 742, the conditional sentence provision, assumes that a conditional sentence of imprisonment is equivalent to prison.

a) Do you think that a conditional sentence of imprisonment is equivalent, or can be made equivalent to a sentence of imprisonment? Why or why not?

b) Does it depend upon the nature of the offence (property; violence; sexual)?
c) The conditional sentence provision also says that it should be consistent with the proportionality principle. Is proportionality relevant when considering a conditional sentence appeal? And if so, when is it relevant?

11. The Criminal Code now suggests that judges should consider non-prison sanctions as much as possible and when appropriate. One way of using prison less is to try to make non-prison sanctions somewhat equivalent.

a) Do you think that a community service order could be made equivalent and substituted for some term of imprisonment? Why or why not?

b) Do you think that a fine could be made equivalent and substituted for some term of imprisonment? Why or why not? Does the offence matter?

12. As mentioned, come have argued that a range of community punishments are easily interchangeable with imprisonment as long as they are severe enough. Would you agree that punishments are easily interchangeable? Why or why not?

13. One of the issues that I'm interested in is public attitudes towards different punishment.

a) Do you think that the public sees a difference between a conditional sentence and probation?

b) Do you think that members of the public will see a conditional sentence as being effective in denouncing crime? Or do you think that they will see imprisonment as more appropriate in accomplishing denunciation?

c) How do you think that the public will react towards this new punishment? Do you think that they will see it as a ‘let off’? Do you think that it depends on the nature of the offence (property; violence; sexual)?
14. I find it interesting to read and listen to how the media describes sentencing. Thus far, do you think that the media is doing a good job, in general, in reporting the use of conditional sentences in sentencing cases? Do you think that they have been accurate?

15. When you think about general problems with conditional sentences, do you think they relate most to instrumental functions of punishment (like incapacitation, or rehabilitation, or individual deterrence), or symbolic characteristics of punishments (denunciation, general deterrence)?
Chapter 6

Multiple Dimensions of Punishment and the Role of the Offence: Implications for Sentencing and Punishment

This thesis has attempted to understand the functions that punishments are seen by members of the public and judges as accomplishing in the context of the purposes of sentencing, the nature of the offence, and the characteristics of the offender. In order to understand the perceived appropriateness of punishment, it is useful to examine dimensions in addition to severity. The results presented in the previous four Chapters have implications for understanding public and judicial sensibilities about punishment, as well as contributing to the literature on punishment and sentencing policy.

The idea that ‘intermediate’ sanctions are generally accepted as substitutes for sentences of imprisonment across all purposes, offences, and offenders is assumed within sentencing legislation for adult and young offenders, as well as by sentencing reformers. The findings presented in the previous four Chapters demonstrate that all ‘intermediate’ sanctions are not equally interchangeable. There are perceived qualitative and quantitative differences among ‘intermediate’ sanctions. Members of the public saw community service orders as more appropriate as substitutes for sentences of imprisonment than fines (Chapter 2). Similarly in Chapters three and five, members of the public and judges did not always view conditional sentences as equivalent and interchangeable with traditional imprisonment. The community service order was also more likely to be seen by members of the public as appropriate for youth than for adults. Fines were viewed as more suitable for adults than for youth (see Chapter two).
The substitution of fines, community service orders, and conditional sentences for sentences of imprisonment was generally more acceptable for property offences as opposed to violent offences. Members of the public were shown in Chapter two to be likely to believe that it was inappropriate to impose a fine or community service order for violent offences. In Chapter four, youth court judges were shown to be more likely to impose an 'intermediate' sanction along with a short period of custody for a violent, sexual or serious property offence than in other types of cases. In the case of conditional sentences, both members of the public (Chapter three) and judges (Chapter five) were willing to impose a conditional sentence for some violence, but not all. Thus interchangeability is not viewed as appropriate across all offences.

One reason why 'intermediate' sanctions are not believed to be appropriate for violent offences relates to the perceived functions of punishment and the purposes of sentencing. Opposition to interchangeability was not related to the perceived severity of punishments. All four Chapters demonstrated that, in varying degrees, denunciation is viewed as an important sentencing purpose to be accomplished through sentencing for offences involving violence.

In Chapter two, it was shown that members of the public did not view fines and community service orders as appropriate for violence because these punishments were seen as lacking the denunciatory power that imprisonment achieves. Those who opposed substitution of these punishments had a belief in the prison – that it denounces crime and more prisons should be built. In Chapter three, opposition to conditional sentences for some violence was shown to be related to a belief that the prison incapacitates and sends a denunciatory message. As demonstrated in Chapter four, youth court judges impose
short periods of custody in order to accomplish denunciation for violent, sexual and serious property offences. 'Intermediate' sanctions alone are viewed as lacking the denunciatory power that a period of custody achieves. Chapter five revealed that judicial perceptions of the unsuitability of a conditional sentence for violence were also related to denunciation. Those judges who believed that denunciation can be achieved by a conditional sentence "almost never and never" -- compared to "always and usually" -- were more likely to mention violence as unsuitable for the punishment.

As an expressive function of punishment, denunciation appears to be important to members of the public and judges. The prison is viewed as punishment -- it is symbolic of what it means to punish. This quality makes it difficult for members of the public and judges to see 'intermediate' sanctions as appropriate substitutes in some circumstances. More broadly, punishment is understood beyond the dimension of severity of the offence or the sanction. The nature of the offence and offender are considered to be important in public and judicial assessments of the appropriateness of punishment. Chapters two to five have demonstrated that it is useful to understand the perceived connections made among functions, purposes, sanctions and offenders.

**Implications for Sentencing Theory and Policy**

Chapter one revealed that on a broad level, an integration of the role of the nature of the offence and offender is a neglected feature of the literature on sentencing and punishment. An examination of some literature on the broad functions of punishment suggested that there is a lack of discussion about how specific punishments are linked to purposes and to the acts that are being punished. More specific analyses about the various
purposes of sentencing also have not focused on how particular sanctions might be more capable than others in accomplishing a purpose of sentencing for a particular offence. Writers on the justifications of punishment have failed to discuss the possibility that some punishments might be seen as necessary and justified for reasons related to the nature of the offence being addressed, and characteristics of the offender (youth or adult).

Authors such as Beattie (1986), Garland (1990), and Garland and Duff (1994) have highlighted the connections among functions, purposes, sanctions, and to a lesser extent, the nature of the offence. These analyses have provided more thorough accounts of how punishment changes over time and place. The present thesis builds on this literature by analyzing the relationships and nuances of the multiple dimensions of punishment in the context of sensibilities toward ‘intermediate’ sanctions and interchangeability.

When conceptualizing interchangeability as a method of reducing the use of imprisonment, it is necessary to consider the purposes that various sanctions are seen as accomplishing for different offences and offenders. As mentioned in Chapter one, approaches to sentencing policy will be inadequate until these dimensions of punishment are addressed and incorporated. The findings presented in Chapters two to five make a contribution to the development of three areas of sentencing policy.

The most obvious issue that arises from this thesis is the challenge to ‘intermediate’ sanctions and interchangeability policies. Clearly it is simplistic to assume that interchangeability will be acceptable across all sanctions, purposes, and offences. Interchangeability guidelines typically offer judges the choice between the “prison” or “non-custodial” punishment, and a range of the length or quantum of the sentence. There
is a lack of opportunity on the part of judges to consider, however, which penalty (other than imprisonment) is appropriate in accomplishing a particular goal of sentencing according to the nature of the offence.

The Minnesota sentencing guideline grid, for example, provides judges with the possibility (at the discretion of the judge) of substituting “non-jail” sanctions along with, or as a substitute for, up to one year in jail (see Morris and Tonry 1990: Figure 2 at 50). All ‘intermediate’ punishments, then, are assumed to be similar and therefore any punishment can be made equivalent to a sentence imprisonment of up to one year. Morris and Tonry (1990) critique the Minnesota guideline grid as an example of the common assumption that “…the only important choice of punishments is between prison and non-prison” (1990: 54). Some authors have attempted to expand the range of possible ‘intermediate’ sanctions in sentencing guidelines. For example, von Hirsch, Wasik and Greene (1992) suggest that different models of paying fines could act as substitutes for sentences of imprisonment (1992: 377).

The research evidence presented here supports these approaches to some extent. In developing a guideline system for interchangeable sanctions it is important to include a specific punishment or range of punishments as possible substitutes for imprisonment. However it also is evident that ‘intermediate’ sanctions are not viewed by the public or judges as operating solely along a dimension of severity, or are seen as all similar. Punishments are perceived to differ both quantitatively and qualitatively. A more effective approach to the development of interchangeability guidelines would be to specify which ‘intermediate’ sanctions are capable of achieving particular purposes of sentencing more than others for different offences.
Another approach to encouraging the use of 'intermediate' sanctions and reducing the use of imprisonment is to provide more specific guidance within present sentencing legislation. Both sentencing legislation for adults and youth include the principle of proportionality without providing more specific guidance for judges on effective ways to accomplish this goal. While the qualitative aspects of the offence have been shown to be important in perceptions of the appropriateness of punishment, the legislation suggests that proportionality in sentencing is determined solely by focusing on severity of the offence and offence history (von Hirsch 1976, 1998). It is likely, however, that the public and judges will assess the proportionality of punishment by considering the appropriateness of the qualities of the punishment, offence, and purposes to be accomplished through sentencing in addition to severity.

One additional way to reinforce the range of 'intermediate' sanctions available to the courts through legislation -- and shift attention away from the prison -- is to create a list of principles that judges must review before determining that the last possible resort is a sentence of imprisonment. In fact Doob (1990) argues that

a statement of purpose and principles is not enough. It may tell judges what principles to follow and may give judges a fairly good idea for a particular case of the appropriate levels of sanction in relation to other cases. But on its own, such a statement does not tell judges explicitly what kinds of sanctions should be imposed for particular kinds of cases....Principles are necessary, but they do not provide sufficient guidance for the sentencing judge (1990: 423).

The incorporation of a list of appropriate purposes that are accomplished by punishments for different offences would create a more effective way of reducing the use of imprisonment and increasing the use of 'intermediate' sanctions.
The second area of policy that may require further consideration relates to the role of the expressive and denunciatory aspects of punishment. It is evident from the findings in this thesis that denunciation is an important feature of public and judicial sensibilities about punishment. However, this purpose is not viewed as easily accomplished by all punishments for all offences, and not easily accomplished by focusing solely on severity of the punishment or offence.\(^1\) This finding may be useful when attempting to develop sentencing legislation.

For example, the proposed Youth Criminal Justice Act (Bill C-3, Second Session, Thirty-sixth Parliament) states that part of the purpose of sentencing young people is to “contribute to the protection of society by holding a young person accountable for an offence…” (section 37 (1)). “Accountability” through sentencing implies responsibility and condemnation of behaviour or denunciation. One principle of sentencing is to achieve proportionality based upon “...the seriousness of the offence and the degree of responsibility of the young person for that offence” (section 37 (2)(c)). There is also a separate list of sanctions available to the courts. This structure implies, to some extent, that any punishment can be crafted to be proportionate to the severity of the offence and also accomplish “accountability” or symbolic goals. Based upon the findings, however, judges will likely require further guidance on which specific penalties are more effective than others in achieving these purposes of sentencing for different offences.

\(^1\) Lacey (1988) reveals the importance of the symbolic and denunciatory aspects of punishment. However, she constructs an argument based upon severity. She maintains that because of the need, in some cases, to fulfill denunciation and disapproval of an offence, it will be necessary to increase the severity of punishments. This discussion does not include an acknowledgment of the qualitative differences among punishments. See pages 193-5.
Related to this discussion is the perception of the denunciatory component of the prison in particular. Chapters two to five demonstrated that it was difficult to accomplish denunciation with fines, community service orders, and conditional sentences for some violent and sexual offences. The finding of the perceived expressive and symbolic quality of prison has been a neglected feature of Feeley and Simon's (1992) analysis of contemporary risk-based approaches to sentencing and punishment (the 'new' penology'). They argue that incapacitating offenders is the main goal of contemporary criminal justice. Clearly the findings presented in this thesis reveal that it is also important to consider in analyses of punishment the denunciatory aspects of sentences of imprisonment, such as the imposition of short periods of custody (see Chapter 3).

Recent discussions of the role of shaming in contemporary punishment relate to a consideration of the expressive functions of punishment. In the United States, for instance, a number of courts are handing out shame-based punishments. For example, those convicted of drunk-driving have been seen wearing derogatory signs -- "I am a convicted drunk driver. And as a result I took a life" -- outside county courthouses (Deardorff 2000). Others have been ordered to place bumper stickers on the back of the car to advertise solicitation of a prostitute. In Canada, there has been a recent newspaper article about the value of shame-based punishments in the United States. The general tone of experts interviewed for the article was negative.

A recent commentary in the United States has considered the positive impact that shaming would have on sentencing practices. Kahan (1999) suggests that

---

The expressive dimension of punishment creates an objective constraint on the opportunities for reforming sentencing practices. This means that substitutes for imprisonment must express comparable condemnation in order to be politically acceptable. If we desire to lessen the severity of the [Federal Sentencing] Guidelines, then, we must find alternatives to prison that meaningfully condemn (1999: 52).

The alternative that Kahan (1999) provides is a shaming penalty for white-collar offenders specifically. He argues that members of the public are not offended by the substitution of shame for imprisonment, such as the use of bumper stickers, signs, and pictures of offenders on billboards and public access television. These shaming penalties are successful, according to Kahan, because they fulfill the expressive dimension of punishment where sanctions such as fines and community service orders fail (1999: 52).

According to Kahan, the use of shaming for white-collar offenders would meet the expressive component of imprisonment and thus an ‘intermediate’ sanction such as a large fine could be attached. Consistent with Federal Sentencing Guidelines, shaming would include ordering the company or individual offender to advertise through the media the nature of the offence committed, the conviction and punishment, and outline steps so that this same offence is not repeated. He argues that the requirement to impose a fine, according to the Guidelines, would be set according to the severity of the offence. In short, this hybrid sentence would decrease the use of imprisonment as well as reinforce shared values in the community.

Kahan’s suggestion of paying attention to the expressive component of punishments when creating substitutes for imprisonment is valuable. It is possible that the use of restrictive conditions for conditional sentences of imprisonment, for instance, may have a positive impact on the perceived denunciatory value of this punishment. However
Kahan does not consider the isolating or disintegrative effects of this method of shaming. First, members of the public will likely believe that the use of shaming penalties only for white-collar offenders is unfair.

Second, as noted in a reply to Kahan by Abramson (1999), a critical factor to be considered is the content of shaming and its implications. Abramson maintains that shaming represents the expression of vindictiveness and exclusion (Abramson 1999: 56). Braithwaite’s Crime, Shame, and Reintegration (1989) argues that shaming ought to be reintegrative rather than disintegrative and repressive. He suggests that family group conferences would be an ideal way of punishing through shaming with the main purpose of reintegrating offenders into the community. His analysis is useful to the extent that he examines issues such as shame and denunciation in more more productive ways than Kahan. Braithwaite’s perspective can have some impact on creating equivalent denunciatory punishments and encourage substitution with sentences of imprisonment.

The findings presented here suggest, then, that punishments express a range of qualities such as punitiveness, loss of liberty, pain, shame, and moral condemnation. Each quality might require individual attention when attempting to create sanctions that are as expressive as the prison and thus be viewed as appropriate and substitutable. Clearly punishment is complex, and sensibilities about punishment should be understood as a number of interacting quantitative and qualitative dimensions.
References


---. "Trends in the Use of Custodial Dispositions for Young Offenders." *Canadian Journal of Criminology.* 34: 75-84.


*R. v. Collins (1987) 33 C.C.C. (3d) 1 (Ont. C.A.).*


*R. v. LaLiberte (2000) SKCA27, 7748 (Sask. C.A.).*


INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.
A Political Anatomy of Detention and Deportation in Canada

by

Anna Pratt

A thesis submitted in conformity with the requirements for the degree of Ph.D.
Graduate Centre of Criminology, University of Toronto

© copyright Anna Pratt (2000)
The author has granted a non-exclusive licence allowing the National Library of Canada to reproduce, loan, distribute or sell copies of this thesis in microform, paper or electronic formats.

The author retains ownership of the copyright in this thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without the author's permission.

L’auteur a accordé une licence non exclusive permettant à la Bibliothèque nationale du Canada de reproduire, prêter, distribuer ou vendre des copies de cette thèse sous la forme de microfiche/film, de reproduction sur papier ou sur format électronique.

L’auteur conserve la propriété du droit d’auteur qui protège cette thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation.

0-612-53690-4
For my mother and father

and

my sweetest Cayenne
A Political Anatomy of Detention and Deportation in Canada
Ph.D 2000
Anna Pratt
Centre of Criminology
University of Toronto

Abstract

This thesis examines exclusionary Canadian immigration law, policy and practices as a distinct penalty: an immigration penalty which polices non-citizens and which administers coercive sovereign sanctions under the authority of the Canadian Immigration Act. It traces the discursive shifts in the governance of this penalty and it attends to the specific practices which these discursive developments have entailed.

The first part of this thesis maps the shifts in the governance of immigration penalty starting with the broadly discretionary, racist and otherwise discriminatory 1952 Immigration Act. The exclusions which it sanctioned were morally-charged, status-oriented exclusions which were governed largely through national purity discourses. It also expressed the continuing influence of Cold War anxieties relating to national security.

Over the 1960s, these guiding rationales were increasingly challenged by liberal legal and humanitarian discourses. The 1976 Immigration Act arguably represents a key point in the extension of liberal legal and humanitarian governance to non-citizens. It 'de-raced' national admission policies, it curbed (some) discretionary powers and it entrenched a permanent on-shore refugee determination system. It legally constructed and confirmed the deservedness of genuine victims of state sanctioned persecution in their countries of origin. However, this triumph of liberal legality and humanitarianism was short-lived. In the 1980s and 1990s, the deservedness of the rights-bearing refugee was increasingly eclipsed by the undeservedness of the free-loading, fraudulent and criminal immigrant and/or refugee claimant. This thesis details the discursive reconstruction of refugees from deserving ‘victim’ to threatening ‘offender’, and considers the legal and policy consequences which this reconstruction entailed.
This analysis considers the steady erosion of the political currency of humanitarian liberal discourses and the rising influence of risk/danger, criminality and victim-focussed discourses in the governance of most areas of public policy. Most relevant here is the way in which anxieties about the threats posed by refugees merged with the discourse on welfare fraud and with broader preoccupations with system integrity which construct the state as victim.

The second part of this thesis examines the implementation of these policy developments. It offers a case study of the processes through which Somali refugees living in Toronto in the mid-1990s were reconstructed as fraudulent, free-loading criminals are examined as are the coercive consequences of this reconstruction. The contemporary preoccupation with criminality that today governs immigration penalty and the proliferation of enforcement-oriented practices and techniques is documented in the second part of this thesis. In the year 2000, Canadian immigration penalty is governed primarily through the rationale of risk/danger, mobilized through criminality, fraud and state-as-victim discourses and operationalised through the interaction of law and discretion. The form of power which is operationalised through this process is sovereign. This dissertation closes with a detailed description of the carceral conditions and penal practices of immigration detention at the Celebrity Budget Inn Immigration Holding Centre in Mississauga, Ontario and with some reflections on the future of detention in Canada.
Acknowledgements

I am deeply grateful to the many people who facilitated and contributed to my doctoral research and to those whose intellectual and emotional support kept me focussed and reasonably sane during even the darkest days of this scholarly journey.

On both a personal and intellectual level, of unparalleled importance to me during the course of this journey was the ever-present and unwavering support of my mother and father, Renate and Cranford Pratt. Their genuine interest in the topic and findings of my doctoral work and the unmitigated joy with which they received even the tiniest of my accomplishments both sustained and inspired me.

I would also like to thank a few of my dearest long-time friends who put up with my long absences and unusual frames of mind over the course of the last five years and who generously helped me in a variety of ways as my thesis and my scholarly self unfolded: Guy Lahaie, El Farouk Khaki, Sian McKenna, Julian Siggers and Henry Knight.

The students and faculty at the Centre of Criminology were always supportive. I will be forever enriched by my intellectual conversations and lasting friendships with my fellow toilers: Kirsten Johnson, Phil Munn, Kim Varma, Voula Marinos, Kimberley White-Mair, Benedict Fischer, Kelly Hannah-Moffat and Willem DeLint.

I have enjoyed the sharp insights and the continuous and always forthcoming support of my supervisor, Mariana Valverde, and the members of my committee, Carolyn Strange and Peter Solomon. I would also like to thank Ursula Franklin, Nicola Lacey and Peter Fitzpatrick.

This thesis would not have been possible without the generous cooperation and contributions of numerous immigration policy officials, refugee lawyers and community advocates in the field. I would like in particular to offer my thanks to: Neil Cochrane and Katherine Clout from Citizenship and Immigration Canada for having taken the time to meet with me and respond to my many queries; Sergeant Ian MacDonald of the Immigration Section of the RCMP for the same reasons; the officials at the Celebrity Inn Immigration Holding Centre in Mississauga who did much more than merely tolerate my
presence there: Fred Franklin of the Toronto Refugee Affairs Committee who facilitated my fieldwork at the Celebrity Inn and whose tireless work on behalf of those in detention and whose commitment to social justice has been both instrumental in terms of helping those in detention as well as inspiring and breathtaking; and to refugee lawyer El Farouk Khaki. whose outstanding advocacy work continues to assist those in need both inside and outside of Celebrity’s walls.

The following people also gave generously of their time to respond to my many questions: Eusavio Garcia of the Quaker Committee for Refugees; Peter Crobbie of the Toronto Family Services Association; Dan D’Ignazio, Adjudicator for the Immigration Appeal Division of the Immigration and Refugee Board; Ali Mohammud, Director of Dejinta Beesha; Constable Rob Purves; Maggie Redmonds, former Community Development Officer for Metro Social Services; and Joan Simalchik, former researcher for the Toronto Refugee Affairs Council Detention Committee.
# Table of Contents

| Title Page | i |
| Dedication | ii |
| Abstract | iii-iv |
| Acknowledgements | v-vi |
| Table of Contents | vii-xvi |

## Chapter One
### Introduction

I) Detention and Deportation as Punishment | 1-3 |

II) The Political Anatomy of Detention and Deportation: An Overview | 3-5 |

III) The Relative Obscurity of Sovereign Power in the Contemporary Context of Immigration Penalty | 6-9 |

IV) Liberal Legality and the Law/Discretion Dichotomy | 9-10 |

V) Rights Versus Risks | 10-12 |

VI) Governing Crime Through Immigration and Governing Immigration Through Crime | 12-15 |

VII) Power and Law | 15-19 |

VIII) Crime as Sign | 19-22 |

IX) The Foundational Exclusions of Liberal Law | 22-24 |

X) Chapter Outline | 24-25 |
# Chapter Two

**Dunking the Doughnut: Law, Discretion and Sovereign Power**

<table>
<thead>
<tr>
<th>Section</th>
<th>Pages</th>
</tr>
</thead>
<tbody>
<tr>
<td>1) Introduction</td>
<td>26-28</td>
</tr>
<tr>
<td>II) The Conventional Legal View of Discretion and the Emergence of Alternative Approaches</td>
<td>28-43</td>
</tr>
<tr>
<td>III) Administration and Judicial Review</td>
<td>43-52</td>
</tr>
<tr>
<td>IV) The Law and Policy Governing the Detention of Non-Citizens Under the Canadian Immigration Act</td>
<td>52-59</td>
</tr>
<tr>
<td>V) Judicial Review of Immigration Decision-Making</td>
<td>59-61</td>
</tr>
<tr>
<td>VI) Departmental Policy Checks on Discretionary Powers</td>
<td>61-65</td>
</tr>
<tr>
<td>VII) Contemporary Policy Debates on Detention</td>
<td>65-71</td>
</tr>
<tr>
<td>VIII) Discretion as Arbitrary and Discriminatory</td>
<td>71-77</td>
</tr>
<tr>
<td>IX) Conclusion</td>
<td>77-79</td>
</tr>
</tbody>
</table>
Chapter Three

Discretionary Power and Discursive Contests: Transitions in the Governance of Exclusionary Canadian Immigration Policy (1947-1976)

I) Introduction 80-83

2) Shifts in the Governance of Exclusionary Canadian Immigration

Law, Policy and Practice (1947-1976): Liberal Challenges and Legal Triumphs 83-116
a) Mackenzie King’s 1947 Statement on Immigration 84-86
b) The 1952 Immigration Act 86-93
   i) Discretionary Power and Exclusionary Provisions 86-90
   ii) Security Concerns 90-93


c) Rising Opposition to the 1952 Act 93-105
   i) The 1962 Immigration Regulations 96-97
   iii) The White Paper of 1966 100-105

   i) The Adoption of the “Point System” (1967) 105-106
   iii) The 1976 Immigration Act 108-116

3) Conclusion 116-119
Chapter Four

Floods, Frauds and Refugees:
The Emergence of the 'Bogus Refugee' in the 1980s

I) Introduction 120-125

II) Reframing Risk in Canadian Immigration and Refugee Law and Policy
in the 1980s 125-133
   i) The Changing International and National Contexts 126-128
   ii) Spontaneous Landings, Public Panics and the Fraudulent Claimant 128-133

III) Stemming the 'Flood': The Administrative Crackdown on 'Bogus' Refugees 133-140
   i) Administrative Measures 133-135
   ii) Legal Reforms: The Role of Bills C-55 and C-84 (1987) 135-140

IV) Expanding the Threat in the 1990s: 'Bogus' Refugee Meets the Terrorist
and the Serious Criminal 140-143

V) In the Name of the 'Truly' Deserving: Bill C-86, the 'Deterrence and
Detention Act' (1992) 144-153

VI) The Creation of the Canadian 'Public Security Portfolio' in 1993 153-154

VII) Conclusion 155
Chapter Five

From Deserving Victims to Undeserving Criminals:
Redefining and Regulating Refugees in the 1990s

I) Introduction 156-159

II) The Rise of Neo-Liberalism as a Rationality of Rule 159-164

III) The Emergence of the Problem of the ‘Welfare Cheat’ in Ontario 164-167

IV) The Merging of Fears About Immigration and Welfare 167-169

V) Extending and Acting Upon the Threat: Cracking Down on the ‘Fraudulent Criminal Claimant’ 170-171

VI) The Provincial and Municipal Dimensions of the Panics Surrounding the ‘Fraudulent Criminal’ Immigrant or Refugee 171-174

VII) From ‘Deserving Victims’ to ‘Fraudulent Criminals’:
The Case of Somali Refugees in Toronto in the 1990s 174-212

i) Somali Refugees: ‘Cheaters’, ‘Criminals’ and ‘Unscrupulous Masters of Confusion’ 177-183

ii) Conflict and Coercion at Dixon 183-191

iii) Fanning the Flames in the Popular Print Media and the Enforcement Response 191-199

a) ‘Dispatch From Dixon’ 192-194

b) Enforcing the ‘Problem’ 194-199

iv) Proof of Identity: Prudence or Prejudice 199-204

v) The Criminalization of Khat 204-212

VIII) Conclusion 212
Chapter Six
Discretion, ‘Danger to the Public’ and ‘Just Desserts’:
The Legal Exclusion of Permanent Residents in the 1990s

I) Introduction 213-214

II) Discretionary Power and Exclusion 214-216

III) Exclusion. Discretionary Power and Bill C-44: A Detailed Interrogation 216-241
   i) The ‘Just Desserts’ Bill: Context 217-219
   ii) Legislative Content of Bill C-44 220
   iii) Administrative Guidelines, Appeals and the Question of Discretion 220-224
   iv) Judicial Review and the Question of Discretion 224-227
   v) Convention Refugees, ‘Refoulement’ and s.70(5):
      ‘Deserving Victims’ or ‘Undesirable Criminals’ 227-232
   vi) Conventional Critiques of s.70(5) 232-241
      a) The Humanitarian Liberal Welfarist Critique:
         The Corruption of Discretion as a ‘Quality of Mercy’ 232-234
      b) The Liberal Legal Critique:
         The Threat of ‘Unfettered’ Ministerial Discretion 234-238
      c) A ‘Radical’ Critique:
         Ministerial Discretion, Exclusion and Racism 238-241

IV) Conclusion 241-242
Chapter Seven

“Criminals First”: Governing Immigration Through Crime in the 1990s

I) Introduction 243-244

II) Administrative Policy Reforms 244-252
   i) ‘Criminals First’ Detention and Deportation Policy (1994) 244-247
   ii) CSIS Joins the Crackdown on Criminals (1996) 247-251
      a) The Expansion of the CSIS Mandate 247-250
      b) Immigration Security Screening 250-251
   iii) Drugs and the Deportation of Refugee Claimants 251-252

III) Creation of Special Departmental Units 252-258
   i) Organized Crime Section of CIC (1994) 252-254
   iii) CSIS Transnational Criminal Activity Unit (1996) 255-257
      a) The Triad ‘Menace’ and Project Sidewinder 256-257

IV) CIC ‘Partnerships’: From Cops to Communities 258-269
   i) The RCMP Immigration and Passport National Enforcement Program 258-264
      a) ‘Community Policing’ and Enforcement 260-263
## The Greater Toronto High Risk Arrest Unit (1995)
- Page 264-266

## Municipal Police Forces
- Page 266-267

## Ongoing Collaborative Enforcement Activities
- **a) CIC-RCMP Taskforce on People Smuggling (1997)**
  - Page 267-268
- **b) Casino Taskforces**
  - Page 268
- **c) CSIS and CIC Airport Collaboration Project**
  - Page 268-269
- **d) Ski-Doo and Boat Patrols**
  - Page 269

## Airlines. Airports and Interdiction
- Page 269-273

## Canada Customs and Crime Control
- **i) Expansion of Coercive Powers (1997)**
  - Page 274-275
- **ii) CANPASS and the ‘Low Risk’ Traveller (1995)**
  - Page 275-276

## Technological Developments
- Page 276-278

## The Production of Knowledge on the Problem of the Criminal Immigrant
- **i) The CIC Metropolis Project**
  - Page 278-281

## ‘Good Guys’ and ‘Bad Guys’: The Entrenchment of an ‘Enforcement Mentality’ Amongst Frontline Immigration Officers
- Page 281-285

## Conclusion
- Page 285-289
Chapter Eight
Detention and Deportation:
Sovereign Power, Carceral Conditions and Penal Practices

I) Sovereign Penality: The Death of Michael Akhimen in Immigration Detention 290-292

II) The 'Celebrity Inn' Immigration Holding Centre 293-328
   i) Carceral Conditions 295-303
      ii) Penal Practices: The Coercive Subjection and Control of Unruly Bodies 303-308
          a) Cuffs and Shackles 304-308
   iii) Coercive Regimes 309-312
       a) Admission Procedures 309-310
       b) Restrictions on Movement 310-311
       c) Patrols, Checks and Searches 311-312
       d) Use of Force 312
       e) Punitive Measures 312-313
   vi) Daily Routines and Schedules 313-319
       a) Time Schedule 313-314
       b) Visitor's Schedule 314-316
       c) Baggage and Valuables Access Schedule 316
       d) Telephone Time Schedule 316-317
       e) Health Services 317-319
   v) Numbers and Composition of Detained Population at Celebrity 319-328
       a) Who is in Detention at Celebrity 320-324
       b) Length of Detention at Celebrity 324-327
       c) Release from Celebrity 327-328

III) The Nature and Scope of Immigration Detention Outside of Celebrity 328-334

IV) Conclusion 334-336
Epilogue: The Future of Detention in Canada 337-343

Sources of Information 344-346

Works Cited 347-363
Chapter One

Introduction

...The traveller glanced casually at the man, who, when pointed at by the officer, had kept his head lowered and now seemed to be all ears, trying to catch something. But the movements of his pressed, pouting lips made it obvious he could understand nothing. The traveller had wanted to put various questions to the officer, but at the sight of the condemned man, asked only: "Does he know his judgement?"

"No," said the officer, about to continue his explanations; but the traveller broke in: He doesn't know his own judgement?"

"No," the officer repeated, pausing for an instant as if demanding a more detailed explanation of the question. The officer then said: "It would be no use informing him. He's going to experience it on his body anyway."

1) Detention and Deportation as Punishment

Michel Foucault observed that in systems of punishment, the body is always at issue: "...the body and its forces, their utility and their docility, their distribution and their submission." He urged that systems of punishment, penalty, cease to be considered as a straightforward means of reducing crime but rather should be analysed as concrete, social phenomena: "...we must situate them in their field of operation, in which the punishment of the crime is not the sole element." The body is therefore enmeshed in a political field: "power relations have an immediate hold upon it: they invest it, mark it, train it, torture it.


3Ibid.. 24
Foucault called on scholars of punishment to consider its "political anatomy". He meant by this that they should attend to the concrete and material political dimensions of the subjugation of human bodies which is accomplished through the discursive processes which turn these bodies into objects of knowledge. As he explained in *Discipline and Punish*:

It is a question of situating the techniques of punishment - whether they seize the body in the ritual of public torture and execution or whether they are addressed to the soul - in the history of this body politic: of considering penal practices less as a consequence of legal theories than as a chapter of political anatomy.  

This study rests centrally upon the proposition that the detention and deportation of non-citizens which is 'administered' by Immigration authorities is, despite repeated legal and political denials, punishment. The conventional legal definition of punishment is "the infliction of some pain, suffering, loss, disability, or other disadvantage on a person by another having legal authority to impose punishment...In modern societies, punishment is generally confined to the criminal law." It is the latter part of this definition that this thesis challenges. The argument that detention and deportation are forms of punishment is not a novel one. Writing in 1959, Victor Navasky argued forcefully that "It is totally unrealistic to deny deportation's penal character" and that "As long as the courts adhere to the fiction that deportation is not punishment, it will be impossible to isolate the breadth and implications of the sovereign power in this area."

In support of this argument, the mention of banishment (and of transportation) is

---

4 Ibid., 25
5 Ibid., 28
common. Banishment is a form of punishment, deportation is a form of banishment. *ergo* deportation is punishment. In 1836, the pain and suffering associated with banishment was eloquently summarised by a critic of the 1798 Alien and Sedition Laws in the United States:

If the banishment of an alien from a country into which he has been invited as the asylum most auspicious to his happiness. - a country where he may have formed the most tender connections; where he may have invested his entire property, and acquired property of the real and permanent, as well as the moveable and temporary kind; where he enjoys, under the laws, a greater share of the blessings of personal security, and personal liberty, than he can elsewhere hope for....if a banishment of this sort be not punishment, and among the most severest of punishments, it will be difficult to imagine a doom to which the name can be applied.  

II) The Political Anatomy of Detention and Deportation: An Overview

This thesis identifies detention and deportation as coercive sanctions, as contemporary manifestations of sovereign power which continue to dominate in the exclusion of undesirable and/or undeserving non-citizens in the 21st century. In the spirit of Foucault, in this thesis exclusionary Canadian immigration law, policy and practices are analysed as forming a distinct penalty: an 'immigration penalty' which polices non-citizens and which administers coercive sovereign sanctions under the authority of the Canadian Immigration Act. It traces the broad outlines of the discursive shifts in the governance of this penalty and it attends to the specific practices which these discursive developments have entailed. It is for these reasons that this study is entitled "A Political Anatomy of Detention and Deportation in Canada".

In the year 2000, Canadian immigration penalty is governed primarily through the rationale of risk/danger, mobilized through criminality, fraud and state-as-victim discourses and operationalised through the interaction of law and discretion. The first part of this thesis traces the discursive shifts in the official governance of national exclusions from the post-second World War period to the mid-1990s in Canada. While, in the most

---

8 quoted in Navesky 1959:221
general of terms, the changes in this governance are described in this thesis as broad ‘shifts’. It should be emphasized that the developments examined in this thesis point to much overlap and hybridity in the governing logics and practices of immigration penalty. In this examination of law and policy, particular attention is paid to the shifting roles of law and discretion in effecting these exclusions. The second part of this thesis focuses more upon the contemporary practices of national exclusions under the current regime of governance, and their impact on those subject to their operations.

As described in Chapter Three, the 1952 Immigration Act was a broadly discretionary, distinctly racist and an otherwise discriminatory piece of legislation. The exclusions which it sanctioned were morally charged, status-oriented exclusions which were governed largely through national purity discourses. It also expressed the emerging influence of Cold War anxieties relating to national security. Over the 1960s, both the national purity rationale as well as the delegation of broad discretionary powers were increasingly contested through the deployment of liberal legal and humanitarian discourses. The 1976 Immigration Act arguably represents a key point in the extension of liberal legal and humanitarian governance to non-citizens. It ‘de-raced’ national admission policies, it curbed (some) discretionary powers and it entrenched a permanent on-shore refugee determination system. It legally constructed and confirmed the deservedness of genuine victims of state sanctioned persecution in their countries of origin. Less widely known though important nonetheless, is the fact that the 1976 Act continued to preserve broad tracts of discretionary power to exclude for reasons of national security and it extended the exclusionary grounds for reasons of criminality.

This triumph of liberal legality and humanitarianism was short-lived. No sooner had the government established a permanent on-shore refugee determination system than it proceeded to limit access to both the shore and the system. The restrictive and enforcement oriented policies of the 1980s and early 1990s were justified by reference to the new and potent threats perceived posed by fraudulent, ‘bogus’ refugees and foreign criminals. The deservedness of the rights-bearing refugee was increasingly eclipsed by the undeservedness of the free-loading, fraudulent and criminal immigrant and/or refugee
claimant. This discursive reconstruction from deserving ‘victim’ to threatening ‘offender’, and the legal and policy consequences which this reconstruction entailed, were effected in part by the steady erosion of the political currency of humanitarian liberal discourses and the rising influence of risk/danger. criminality and victim-focussed in the governance of most areas of public policy. Most relevant here is the way in which anxieties about the threats posed by refugees merged with the discourse on welfare fraud and with broader preoccupations with ‘system integrity’. In this reconstruction, the state. ironically, emerges as the ‘victim’.

In the second part of this thesis, these policy developments are examined in their practical application. The processes through which Somali refugees living in Toronto in the mid-1990s were reconstructed as fraudulent, free-loading criminals are examined, as are the coercive consequences of this reconstruction. While the particular situation of Somalis living in Canada is the only mini-case study provided, it would be most interesting to examine the ways in which these developments are experienced by other groups of new immigrants and refugees in Canada.

The contemporary preoccupation with criminality that today governs immigration penalty to a degree unprecedented in Canadian history, and the myriad enforcement-oriented practices and techniques which have proliferated under this rationale, are documented and detailed in the second part of this thesis. This dissertation closes with a detailed description of the carceral conditions and penal practices of immigration detention: of the operation of sovereign power in Canada’s own ‘Kafka motel’: the ‘Celebrity Budget Inn’ in Mississauga, Ontario which is also an Immigration Holding Centre.⁹

---

⁹Airport motels in the United States which are used for immigration detention are commonly, and indeed appropriately, referred to as ‘Kafka motels’.
III) The Relative Obscurity of Sovereign Power in the Contemporary Context of Immigration Penalty

....How different the executions used to be! The entire valley was mobbed a whole day in advance; they all came just to watch. Early in the morning, the commander showed up with his ladies; fanfares awoke the entire camp;...The machine was freshly polished and glistening; I used new replacement parts for almost every execution. In front of a hundred eyes - all the spectators stood on tiptoes as far as the hills - the commander himself placed the condemned man under the harrow...Everyone knew: Justice was being done. In the hush, all we heard was the condemned man's sighs, muffled by the gag...It was impossible to allow everyone who requested it to watch from up close. The commander, wise as he was, ordered that preference be given to the children. I, however, because of my profession, was always permitted to stand close by; I would often squat there with a small child in each arm, right and left. How profoundly we took in the transfigured expression from the tortured face, how intensely our cheeks basked in the glow of that justice, attained long last and already fading! What wonderful times, my friend!”

Executions and the range of other bodily mutilations characteristic of pre-modern and early modern penalty no longer take place in the public square. That sovereign power continues to reign supreme over the exclusion of undesirable and undeserving non-citizens in the 21st century and that it does so in relative obscurity and secrecy is of significant analytical interest. In addition to the necessary corporeality of sovereign punishments, as described by Foucault, the effectiveness of sovereign penalty well into the early Modern period relied upon the visibility and public nature of its bodily sanctions: the ‘spectacle’ and ‘ceremony’ of punishment. Foucault documented how the emergence of a ‘higher aim’ of punishment, that of correction, served to displace the body “...as the major target of penal repression.”" The disappearance of torture and public square executions marked “...the decline of the spectacle... (and) a slackening of the hold

10 Kafka. “In the Penal Colony” 1995:209-211

11 Foucault. Discipline and Punish. 1979:8
on the body. Foucault conceded that a "trace" of torture persists in the modern mechanisms of criminal justice: "...a trace that has not been entirely overcome, but which is enveloped, increasingly, by the non-corporal nature of the penal system."

The operation of sovereign power in the exclusion and expulsion of undesirable and undeserving non-citizens in the current context of immigration penality is thus markedly different from the days in which offenders were ceremoniously tortured in the public square. However, in contrast to modern criminal justice penality, the "body" of the undesirable non-citizen remains the primary target of the sanctions imposed. Unlike criminal justice penality, there is no "higher aim" at work in the detention and deportation of non-citizens: there is no discipline. Detention and deportation and the various techniques of bodily subjection which they entail, are not merely "traces" of the bodily preoccupations of the sovereign power of days long gone by, but rather represent the continuing centrality of the operation of sovereign power in the exclusion and expulsion of undesirable and undeserving non-citizens. Just as the public execution of the past was "...much more than an act of justice: it was a manifestation of force: or rather, it was justice as the physical, material and awesome force of the sovereign deployed there..." the same may be said of detention and deportation.

That today this power operates behind closed doors instead of in the public square is indicative of the degree to which it has been reconfigured by the influences and preoccupations of liberal legality. However, in the context of immigration penality, it is "traces" of liberal legality that inflect the administration of the bodily sovereign sanctions of detention and deportation rather than the reverse. While the painful and public excesses of the administration of sovereign power have been circumscribed, the inherent violence of these fundamentally bodily sanctions continues to be their defining feature. The essential violence of detention and deportation is not a "barbaric hangover", but as

---

12Ibid. 11
13Ibid. 16
14Ibid. 50
Foucault observed in the context of criminal punishment. "...is actually inscribed in the political functioning of the penal system."\(^\text{15}\) Nevertheless, the spectacle of these sovereign sanctions is clearly much diminished in the present day. Immigration penalty is even more secret and mysterious than criminal justice penalty: it is more withdrawn from the public eye and receives far less critical scrutiny than do prisons. Moreover, those subject to detention and deportation proceedings have fewer rights than prison inmates, not only legal rights but also everyday rights such as fresh air, exercise and direct contact with friends and family members.

In addition to the fact that liberal legality has dulled the sharp edge of the sovereign's sword in this area, it is perhaps also true that unlike public executions, the sanctions of detention and deportation need not strike terror into the hearts and minds of all people: their spectacular deterrent effect is particular to non-citizens. Nevertheless, there is still a degree, albeit limited, of spectacle involved. Word of mouth, occasional sensational media accounts of detentions and deportations, and political rhetoric all contribute to their continuing symbolic power. The lives and activities of non-citizens in Canada are indeed governed, in a very real way, by the 'terror' inspired by detention and deportation.

...I know its impossible to make those times comprehensible today. Anyway, the machine is still running and it still works on its own. It works even when it's alone in this valley. And, ultimately, the corpse, in an incomprehensibly gentle flight, still drops into the pit even if hundreds of people no longer gather around the pit like flies, as they used to do. Back then, we had to install a sturdy railing around the pit, but it was torn down long ago.\(^\text{16}\)

**IV) Liberal Legality and the Law/Discretion Dichotomy**

\(^\text{15}\)Ibid., 49

\(^\text{16}\)Kafka ""In the Penal Colony"" 1995: 211
National exclusions are accomplished through the interaction of law and discretion. This interaction facilitates the operation of sovereign power against undesirable and undeserving non-citizens. The conventional view of law and discretion, as an oppositional binary, serves to further obscure and distract attention away from the violence and coercion of detention and deportation, by setting and limiting the parameters of interrogation. That is to say, liberal legality abstracts and redefines real problems in legal terms. Coercion, discrimination and other abuses associated with national exclusions are reconceptualized as exceptional problems of 'discretion', usually framed in terms of 'arbitrariness'. The largely taken-for-granted solution to these perceived problems of discretion is more law or law-like rules. Nicola Lacey offers the following definition of the legal paradigm:

...the subjection of areas of human conduct and practice to regulation according to clear, prospective, publicly announced general rules or rule-like standards. Problems are typically seen as arising from ambiguities or 'gaps' in the rules, calling for clearer interpretations or further legislation or quasi-legal action. Disputes are seen as calling for resolution on the basis of given rules and according to standards of due process. This approach is closely associated with the ideal of the 'rule of law' and hence with liberalism as a doctrine of political morality.\(^1\)

In Chapter Two, I provide a critical overview of the discursive parameters of the scholarly debates on law/discretion and consider the policy implications of the dominant discourse on discretion in the specific context of Canadian Immigration detention and deportation decision-making. In this analysis, I propose that the limits imposed by the conventional dichotomous view of law/discretion may be avoided by considering discretion as an activity, a technology employed in the governance of national exclusions and of 'undesirable' and/or 'undeserving' non-citizens. Chapter Three considers the historically shifting roles of law and discretion and the different political rationales and


\(^{18}\)Ibid..362
discursive contests that have governed their operation and inflected their use in Canada.

V) Rights Versus Risks

In Chapters Four and Five of this thesis, I seek to establish that the liberal legal conception of refugees as ‘rights bearing’ individuals and the related liberal humanitarian conception of refugees as ‘deserving victims’ have been steadily unsettled and increasingly supplanted by a view of refugee claimants (and indeed new immigrants in general) as posing various risks to the nation and public. This reconstruction of refugees from being ‘at risk’ to being inherently ‘risky’, evidences the increased prevalence of ‘risk’ as a rationale of governance. This development is consistent with the general observations made in the emerging criminological literature on risk (‘new penology’), about the emergence of a ‘risk-based’ society in which individuals and populations are increasingly governed through techniques of risk management.19

While it is certainly true that risk discourses have indeed become more prevalent in the governance of national exclusions, the degree to which this development has entailed the real changes in either the actual technologies of governing or in the political

---

and moral underpinnings of the categories of risk is far less certain. Further, as pointed out by Kelly Hannah-Moffat in the context of women's imprisonment in Canada, "...the presumption that risk governance acts uniformly across whole populations rather than differently according to gender, race and other variables..." (to which I would add citizenship status), must be critically examined.20

Mary Douglas theorises the vast discrepancy between, on the one hand, the 'neutral and objective' origins of risk theory and probability calculations and, on the other, the political and public currency and deployment of the concept of risk. Her insights are particularly resonant in the context of national exclusions. Douglas attends to the politicization of the concept of risk. She observes that while the view of risk assessment as a scientific, "... purely neutral, objective tool of analysis", vests the language of risk with its political power. In reality, the concept of risk and its persuasive power has very little to do with science: "...the risk that is a central concept for our policy debates has not got much to do with probability calculations. The original connection is only indicated by arm-waving in the direction of science. The word risk now means danger: high risk means a lot of danger." 21 She explains,

The language of risk is reserved as a specialized lexical register for political talk about undesirable outcomes. The charge of causing risk is a stick to beat authority, to make lazy bureaucrats sit up, to exact restitution for victims. For those purposes danger would once have been the right word, but plain danger does not have the aura of science or afford the pretension of a possible precise calculation.22

Douglas observes that despite its modernist origins, the political work of the concept of risk/danger resembles "antique" preoccupations with sin and taboo. Both construct a 'moral community' in need of protection. Risk/danger, like sin, is used to

20 Kelly Hannah-Moffat, "Moral Agent or Actuarial Subject: Risk and Canadian Women's Imprisonment". (Draft) January 27, 1998


22 Ibid., 24-25
"...protect individuals from predatory institutions or to protect institutions from predatory individuals." In a manner reminiscent of Durkheim on social solidarity, Douglas observes that "It may be a general trait of society that fear of danger tends to strengthen lines of division in a community." In the context of national exclusions, Douglas' analysis of risk/danger is most appropriate. While it is indeed true that risk/danger occupies an increasingly prominent place in the governance of national exclusions and while it has underpinned a wide range of enforcement oriented initiatives, there is little evidence that this development has entailed the use of new or different 'risk-assessment' techniques or that it has resulted in any significant changes in the populations being excluded on the basis of risk/danger.

When viewed in this light, risk/danger emerges as a powerful rhetorical exclusionary technique and a central mechanism for the related processes of constructing collectivities including, of course, that of the nation. In the context of immigration exclusions and the development of law and policy in this field, risk means danger, and conceptions of danger continue to be constituted along the lines of race, morality and increasingly 'nationality'. As such, the concept of 'risk' is easily deployed as "... a slogan for mustering xenophobia."  

VI) Governing Crime Through Immigration and Governing Citizenship Through Crime

In the current context of immigration exclusions, the rationale of risk/danger is constituted through and mobilized by crime and criminality discourses to an extent unprecedented in Canadian history. As I examine in Chapters Four and Five, dominant conceptions of criminality have come to include (certain kinds of) fraud: the 'foreign criminal', the 'welfare cheat' and the 'bogus refugee' have been increasingly linked and

---

23 Ibid. 26
24 Ibid. 34
25 Ibid. 39
constructed and acted upon as risky 'predators' who pose a serious threat to the Canadian nation and public. Over the last two decades, exclusionary, restrictive and enforcement-oriented law, policy and practices have proliferated in the official effort to crack down on (reduce the 'risks' posed by) criminal immigrants and refugees. Chapters Six and Seven document the existence of an expansive and coercive immigration penalty which is justified primarily through reference to the risks/dangers posed by foreign criminals. On the level of public policy, exclusionary immigration penalty has emerged alongside, and intertwined with, law enforcement agencies as a central and largely taken for granted mechanism for the policing and punishment of non-citizens in Canada. It is in this sense that crime is today governed, in part, through immigration.

However, this official preoccupation with governing crime through immigration, and the range of crime control initiatives and practices which this preoccupation has entailed, serves to govern much more than crime. Indeed, the relative absence of any compelling evidence regarding the size of the crime problem posed by foreign-born criminals in Canada, as well as the relatively small numbers of people who are directly subject to coercive national exclusions on this basis, as well as the particular racial, ethnic and national characteristics of those being subjected, suggest that some other work is being done: something other than crime is being 'governed through crime'.

Jonathan Simon examines the emergence of this new mode of governance in the United States. Drawing on Foucault, he explains that "...We govern through crime to the extent to which crime and punishment become the occasions and the institutional contexts in which we undertake to guide the conduct of others (or even of ourselves)." Simon observes that the actual nature of the quantitative increase in crime in the United States cannot alone explain the trend toward governing through crime. And while the influence of the "..resurgence of conservative political forces...[which] favour the..."

---


27Ibid.,174
criminal law as a tool of state governance" has been considerable. Simon notes that conservatives are not alone in their quick recourse to law and punishment based governance: liberals too "...are drawn toward punishment as the locus for governance." 28

Simon argues that the trend toward governing through crime is part of a more general crisis in governance ushered in by the decline of welfarist liberal modes of governance: that the contemporary governmental preoccupation with crime and punishment is consistent with new neo-liberal technologies of governing which emphasise the importance of individual (as opposed to collective, socialised) risk management. Simon observes that governing through crime "is ...a way of imposing this model of governance on the population." 29 While this may be the case for citizens. in the context of national exclusions. non-citizens deemed undesirable and/or undeserving through the operation of these criminality discourses are not incited to better govern their risks: they are simply detained and deported.

Simon suggests that this trend toward governing through crime is particularly heightened in the United States because there crime and punishment have been long been intertwined with the problem of race:

...the real and imaginary links of violence...with young African American men are helping to drive the imperative of governing through crime. Whether or not voters acknowledge such motives to pollsters. it is hard to ignore the continuities between the present situation and a traditional preference for governing predominantly African-American populations in distinct and distinctly less respectable ways. 30

While the racist and discriminatory treatment of aboriginals in Canada by state authorities. including criminal justice authorities. has a long and entrenched legacy in Canada. arguably. the development of Canadian criminal justice penalty has not been as severely inflected by race as in the United States. However. the development of Canadian

28Ibid..176
29Ibid..178
30Ibid..181
immigration penalty certainly has. Race has always been a key variable in the governance of exclusionary Canadian immigration policy and new immigrants and refugees. The mobilization of crime and criminality in the governance of national exclusions produces and reproduces dominant constructions of the (un)desirable Canadian citizen. Historically specific constructions of the threatening outsider. the undesirable non-citizen. the ‘other’. are constituted. in varying degrees. along the lines of race. ethnicity and more recently nationality. For those non-citizens constructed and acted upon as undesirable. the dominant mode of power to which they are subjected to is coercive and repressive sovereign power.

The historical legacy of explicitly racist and otherwise discriminatory Canadian law and policy and the less explicit forms of racism which persist today (including. for example. the dearth of Canadian immigration visa offices in non-white. poor countries and the systemic discrimination which continues to inflect admissions criteria). speaks to the existence and persistence of race as a constitutive element of dominant conceptions of the (un)desirable non-citizen. The fact that the vast majority of non-citizens who are arrested. detained and deported are non-white. that the majority of those deemed to be a ‘danger to the public’ are non-white. and that the official criminal profiles used in the policing of non-citizens are explicitly organized according to race (ethnicity and/or ‘nationality’) are indeed. in Simon’s careful words. “hard to ignore”.

VII) Power and Law

The violent exclusion of ‘undesirable’ and ‘undeserving’ non-citizens through detention and deportation. and the discursive processes and material practices which inflect and accomplish this subjection. necessarily raise provocative and intellectually productive questions about the nature of governing power and the role of law under a liberal regime of governance. Two characteristics make this field of governance distinctive: the fact that those being coercively governed are not citizens and that the mode of power to which they are subject is distinctly sovereign.

The detention and deportation of undesirable and/or undeserving non-citizens
raises the issue of the persistence of ‘illiberal’ modes of power into the 21st century.\textsuperscript{31} Also evident in this study is the continuing primary place of territorial sovereign law and coercive sovereign penalty in the governance of national exclusions. These exclusionary modes of governance are ‘state-centred’; they continue to depend upon the legitimacy and coercive force derived from the sovereign state, as it is conventionally understood.

The national border continues to be grandly and practically defended through the deployment of sovereign rationales which underpin the actual practices of ‘controlling borders’. Entering Canada is a ‘privilege’ and not a ‘right’ and as such the infringement of liberal liberties at the nation’s borders is legitimised. The practices employed in the work of controlling the borders by customs and immigration officials continues to be justified explicitly and unapologetically by reference to sovereignty. Customs officers and front-line immigration officers carry out moment to moment infringements of liberal liberties in the name and interest of sovereignty. It should be emphasized that sovereign power at the nation’s borders governs citizens as well, although the effects of the operation of sovereign power are more severe for non-citizens.

Of course, sovereignty is not the only way that non-citizens (or indeed citizens) are governed. Indeed, Canada is arguably one of the most thoroughly liberal of western liberal democracies in respect to non-citizens. Indeed, the extension of the legal protections afforded under the Canadian Charter of Rights and Freedoms to all people on Canadian soil\textsuperscript{32}, citizens and non-citizens alike, evidences the breadth of this liberal legal commitment as does the legal status afforded to permanent residents living in Canada. *Canada is not a police state*. Indeed the degree to which governance has been legalized, including most notably for the present purposes, the governance of exclusionary immigration law and policy, is breath-taking. It is nonetheless true that in the ‘defence’ of


\textsuperscript{32}As confirmed in the famous 1985 *Singh* decision.
the nation’s borders, sovereign rationales and sovereign power trump all other considerations.

However, territorial sovereign law (understood as a straightforward expression of the will of the sovereign) certainly cannot explain the particular and specific policies and practices which are generated and applied in the field of national exclusions. Even the application of coercive sovereign power in the exclusion of undesirable and undeserving non-citizens must, under a liberal regime of government, be rationalised and legitimised by something other than sovereignty. In response to Alan Hunt’s call for scholars to “re-open the problem of sovereignty” and re-acknowledge the discursive power of juridical law,33 Marianne Constable has evidenced the coexistence of different modes of power in the governance of immigration in the United States through a detailed analysis of two legal texts. Constable pays particular attention to the rise of governmental preoccupations in the justification and legitimization of American immigration law. She argues that while sovereignty continues to inflect the justifications for immigration legislation, what Foucault calls governmental preoccupations should not be ignored. Drawing attention to the interplay between different modes of governance evident in the texts of American immigration legislation, she argues that the non-sovereign rationales deployed are largely ‘governmental’: i.e. pertaining to the national interests of the population:

Even as the modern state carries out its ‘sovereign’ powers and even as sovereignty is extended to and appropriated by the collectivity, the exercise and articulation of such power reveals the presence of a “governmental rationality”, which justifies policies...by appealing to the complex interests of the population. Instead of grounding law in divine law or nature, modern policies - both state and non-state - for the distribution of goods ground themselves in knowledges of populations produced by social sciences.34


Constable argues that the governmental rationale of "managing resources" is a central feature of American immigration law and policy. Following Foucault, Constable aims to reveal the interpenetration of different governing rationales in the context of American immigration law (the "sovereignty-discipline-government triangle") as well as to clarify and reassert Foucault's argument regarding the analytical limits of a juridical conception of sovereignty which treats sovereignty and law "...as absolutely one and the same thing." In response to Hunt's concerns about Foucault's "expulsion" of law, Constable counters that: "Foucault does not "repulse" or "expel" law or the state from modern analyses of power; rather, he denies the capacity of sovereignty - as power and as theory - and of law conceived of as sovereign to account for the rationality of the modern state."

There is little question that immigration law articulates and employs different modes of governance, and that sovereignty is but one of these rationales. Nor is there any doubt that governmental concerns with the national interest are central justifications for specific immigration policies. Nonetheless, the relative importance and centrality of sovereign power cannot be discerned by looking at legal texts alone. As the present study documents, the actual practices of controlling borders and effecting national exclusions remain expansively and unapologetically sovereign. Moreover, from the perspective of those subjected by and excluded through the practical operations of sovereign power governmental rationales are of little importance.

It is suggested here that while it is certainly true that different forms of power certainly co-exist in the governance of individuals and populations, it is a distinctly coercive and sovereign mode of governance which operates on those who fall outside of the definitional boundaries of the juridical citizen-subject. In the case of non-citizens, these boundaries are explicit. The violent exclusion of undesirable and undeserving non-

---

35 Ibid., 256
36 Ibid., 252
37 Ibid., 253
citizens powerfully exemplifies the illiberal treatment of those who both literally and symbolically fall outside of these legal and territorial boundaries. While governmental justifications are certainly deployed in support of national exclusions, no governmental efforts are actually needed to transform these 'deficient' outsiders into desirable citizens: when non-citizens are deemed undesirable, they are excluded and/or expelled.

VIII) Crime as Sign

*Founded upon the illusion of the social contract, [the nation's] inclusionary 'we'...

...turns out to be exclusionary.

*The 'we' can only exist through the expulsion of the Others.*

This study points to the ways in which the coercive national exclusions work, not only to repress and exclude, but to *produce* dominant constructions of the desirable and deserving citizen and to define and enforce national boundaries - boundaries which are not merely sovereign and territorial, but which rest centrally upon moral conceptions of desirability and deservedness. Not all non-citizens are excluded; some possess attributes which facilitate their inclusion. An analytical preoccupation with identifying different rationales of governance in the abstract does little to illuminate their moral dimensions or the ways in which their applications are differently distributed along the lines of race, ethnicity, nationality, citizenship, gender, class, etc.

The moral dimension of national exclusions hinges upon the distinction between 'deserving' and 'undeserving', 'desirable' and 'undesirable'. As with the distinction between the deserving and undeserving poor in the context of social welfare provision, the distinctions between the deserving and undeserving refugee, and the desirable and undesirable immigrant have the dual consequence of moralising the categories and legitimising differential treatment according to these moralised categories.*9 The moral

---


element of these categories "...involves any normative judgement that some conduct is intrinsically bad, wrong or immoral. It is an important supplement that moralising discourses frequently invoke some utilitarian consideration linking the immoral practice to some form of harm."40 As such, this "dividing practice"41 plays a key role in the production and reproduction of borders and boundaries, including national ones.

A central technology in the operationalisation of this exclusionary "dividing practice" is "crime". The operation of "crime as sign" in the construction and enforcement of, in this case, national borders and boundaries provides a valuable insight into the productive power of crime and victim discourses and the ways in which the construction of crime and criminality is inflected by such categories as race, gender and class.

As explained by Alison Young, the processes of criminalization and victimization are related in dichotomous opposition and work together to produce "the community" and different representations thereof. In this view, this production of communities requires the continuous and repeated "sacrifice" of the threatening outsider, the "outlaw". Young's observations regarding this process in the particular context of the nation are particularly relevant to this study:

The construction of a community founded upon the expulsion of textual outlaws ... requires the establishment of borders and boundaries. Beyond the boundary can be identified the outlaw. within the community exist the members of the community attempting to "lead their lives free from fear and in relative security"... The localism of crime necessitates the nomadism of the law's response. ...Prediction becomes a major bulwark of the community's fight against crime: for prediction can suggest future locations of the boundary... Locating crime, pinning it down, becomes a major preoccupation of the crimino-legal complex. Where crime is located, the border can be strengthened. As crime is deemed continually to take new forms, new borders come into crisis. requiring reinforcement and vigilance...

Some borders are easily identified: immigration law establishes the nation...

40 Ibid. 7
41 Ibid. 8
as a discrete community which is vulnerable to flooding by individuals from other countries. The recent development of a specific form of immigration policing... testifies to the increasing investment in and anxiety about the (in)vulnerability of the nation. Recent years have seen the intersection of criminal law, immigration policy and the criminal justice system in dealing with immigrants through detention and deportation.42

Governing through crime is accompanied by its corollary, governing through victimization, and victimhood has increasing become a new basis for inclusionary claims to citizenship:

Crime control becomes a matter of minimizing the risk of victimization.
...Passivity, for the victim, is initially unavoidable in that crime, like illness, happens to the individual, with all the force and randomness of circumstance. Agency can be regained. ...if the individual rejects such passivity and takes up a role in the prevention of crime. In this way, citizenship is acquired. As an active social agent, the citizen will fit locks and bolts to doors and windows, avoid dark streets and purchase alarm systems. If everyone is a victim, than everyone has a part to play in the struggle against crime.43

In the context of the formation, defence and reproduction of national borders and boundaries, it is the nation-state (and by extension the nation’s 'public') that is constructed as the primary victim of the foreign crime problem and it is the state, as victim, that must be vigilant and proactive in the struggle against crime and criminals.

A stark example of the deployment of criminality and victim discourses, and the 'dividing practice' which pits the 'deserving' against the 'undeserving', in the construction and fortification of national boundaries through restrictive immigration law and policy is provided by a 1997 issue of the International Journal, published by the Canadian Institute of International Affairs. The cover of this issue promised two articles on "The New Slave Trade". The first, written by journalist Daniel Stoffman, bears the title "Making Room for Real Refugees" and the second, written by former Immigration

42Young. Imagining Crime. 1996:20

43Ibid..56
and Refugee Board Member and Canadian diplomat William Bauer is entitled “Refugees. Victims or Killers”. 44

Historically specific constructions of crime and criminality are, to varying degrees, necessarily inflected by race, as they are by gender, class, nationality, and morality: “[T]he body of crime is being continually reconfigured as feminine, black, young, homosexual, maternal and on and on. Such a process does not and cannot end.” 45 It is through these never-ending productive processes of exclusion that these variables are translated into national policies and practices of exclusion.

IX) The Foundational Exclusions of Liberal Law

In the continuing efforts to theorize race and racism without resort to conventional structural analyses, some scholars have located a foundational exclusionary dynamic within liberal law itself. Peter Fitzpatrick has considered the dynamics of negation in the context of liberal law. However, unlike scholars like Young who concentrate on ‘exclusion’, Fitzpatrick seeks to understand both the inclusions and exclusions of law. In his view, the constituent force of liberal law is the constant effort to both negate and include: the outsider is called on to be the same yet at the same time is constructed as different and excluded. This dynamic issues out of what Fitzpatrick terms the ‘mythology’ of modern law.

Liberal law, while appealing to transcendence and universality, is necessarily historically specific and particular. Fitzpatrick proposes that this central contradiction (between the universal and the particular) can only be reconciled by considering the transcendence of law as myth and the ways in which the mythology of modern law is continuously constituted through negations;

Thus modern law emerges, in a negative exaltation, as universal in opposition to the particular, as unified in opposition to the diverse, as omniscient in contrast to the incompetent, and as controlling of what


45 Young. Imagining Crime 1996:19
has to be controlled...Law is imbued with this negative transcendence in its own myth of origins where it is imperiously set against certain "others" who concentrate the qualities it opposes.46

While law claims universality, it simultaneously effects closure: it is a "fixed point". The "fixed point" of law is constituted in part through the opposition of savagery and civility. Fitzpatrick argues that the inclusionary liberal legal principles of equality and universality, which stand in opposition to racism, actually work "import racism into law".47 As observed by Fitzpatrick, the western liberal subject and liberal legality are constituted through the negation of the "other", the myth-ridden and uncivilized savage. In this view, savagery is an ever present feature of liberal law: it is outside of law, but always proximate to law. In this way racism is necessarily a key term in the operations of liberal law.48

One final insight drawn from Fitzpatrick's analysis of liberal law is important to this thesis. Discretion is a key component in the reconciliation of the universality and particularity of liberal law. In this view, discretion is erected as an oppositional barrier to fixity: if fixity doesn't exist, then either does discretion. The discretion/law dichotomy is


a thus a central constitutive feature of the mythology of modern law. Fitzpatrick frames this issue in terms of what he refers to as the mythic mutuality of law and administration: "[I]t is because of the particularist and pervasive powers of administration that the rule of law can be maintained in all its aspects of universality and equality and seen as marking out fields for free action...Administration is the necessary 'dark side' of law."49 This thesis incorporates this insight by contributing to the deconstruction of the law/discretion dichotomy and by considering the practical role of discretionary power in accomplishing national exclusions.

X) Chapter Outline

Chapter Two provides an overview of the contemporary literature relating to law/discretion and considers the limitations imposed by the dichotomous opposition of law and discretion in the context of Canadian Immigration detention decision-making policy and practices. It proposes that the analytical limits imposed by this dichotomy can be avoided by considering discretion as a form of power. Chapter Three provides an overview of the shifts in the governance of exclusionary Immigration law and policy from the 1952 Immigration Act up to, and including, the 1976 Immigration Act. It pays particular attention the role of discretion and the rise of criminality concerns in national exclusions. Chapter Four examines the rise of 'risk-thinking' in the particular context of refugee law and policy in the 1980s and details the reconstruction of the 'deserving refugee' to the 'undeserving' and inherently 'risky' refugee claimant. Chapter Five documents the emergence of the rationale of risk/danger as the dominant mode of governing immigration penalty. This rationale, constituted primarily through crime and criminality discourses, is seen to merge with broader neo-liberal concerns with 'system integrity' and welfare fraud. The concrete effects of these developments are examined through a case study of the reconstruction and regulation of Somali refugees living in Toronto in the mid-1990s.

Chapter Six examines the 1995 extension of ministerial discretion to exclude permanent residents deemed to be a ‘danger to the public’ under Bill C-44. This piece of legislation is held to represent the political and legal manifestation of a powerful shift in the governance of Immigration penalty in which ‘criminality’ concerns have merged with ‘security’ concerns and have emerged as the dominant rationale of governance in this field. Chapter Seven documents the emergence of exclusionary Immigration law, policy and practices as a central and largely taken-for-granted mechanism of national crime control. It details the range of enforcement-oriented developments and initiatives which have proliferated in the official efforts to crack down on the ‘problem’ of the criminal immigrant. Chapter Eight describes and examines the carceral conditions and penal practices of Immigration detention at Celebrity Inn, the Immigration Holding Centre in Mississauga. This study closes with an epilogue which reflects upon the future of immigration detention in Canada.
Chapter Two

Dunking the Doughnut:
Law, Discretion and Sovereign Power

Discretion, like the hole in a doughnut, does not exist except as an area left open by a surrounding belt of restriction.\(^1\)

I) Introduction

This often cited ‘doughnut analogy’ of Ronald Dworkin neatly encapsulates the conventional view of discretion. Three main assumptions are embedded in this view - that law is the primary instrument of social regulation, that discretion is a residual category of law,\(^1\) and that this discretion is exercised by individuals who, though influenced in a wide variety of ways, are essentially autonomous. While recent scholarly analyses of discretion have begun to unsettle this conventional view, its core assumptions nonetheless continue to underpin political discussions and policy debates to such an extent that it sets the parameters of imagined policy and legislative reforms related to its use. This is the case despite contemporary and historical experiences which cast some doubt on the potential of law to effectively address the ‘problem’ of discretion.

This chapter aims to contribute to the unsettling of this conventional view of discretion and introduces an alternative understanding of discretionary power, the understanding which informs the this thesis as a whole. It begins with a critical review of the academic literature on discretion, paying particular attention to the liberal assumptions identified above. Reflecting a belief in the importance of contextual analysis, it then discusses the uses of discretion in the context of the administration of a body of law in which discretion is particularly important but has not been much examined - the enforcement provisions of the Canadian \textit{Immigration Act} (forthwith referred to as the Act), particularly with respect to the detention, release and removal provisions. Current legal, activist and political debates around discretion and detention are then discussed in

\(^{1}\)Ronald Dworkin “Judicial Discretion” 60 \textit{J. Philosophy} 1963:624
the light of the 1997-1998 legislative review on Canadian immigration law and policy in
general and of the subsequent parliamentary hearings on the specific issues of
immigration detention and removal. The chapter ends with some reflections on how one
might avoid, conceptually and methodologically, the conventional view of discretion and
the limited legislative and policy reforms which an acceptance of this view entail.

Discretion has long been the focus of scholarly and policy related discussion and
analysis in the field of criminal justice. Much attention has been paid to the uses of the
discretionary powers of judges, prosecutors and the police. Surprisingly, despite the
obvious comparisons between the enforcement powers related to the Act and those related
to the Criminal Code, little analytical attention has been paid to the wide scope of
discretionary powers accorded to immigration officials, particularly with respect to their
broad powers of arrest, detention and deportation.

This is due, in significant part, to the analytical preoccupation with judicial
settings that issues out of the dominant liberal legal paradigm and to the acceptance that
the rights normally central to that paradigm are not applicable to non-citizens seeking
admission to Canada. Clearly, whenever the law delegates discretionary powers which
permit the abrogation of fundamental liberal rights and freedoms the issue is of particular
concern. When this abrogation entails the application of extreme and coercive bodily
sovereign sanctions, as is the case with immigration detention and deportation decision-
making, the issue of discretion becomes that much more compelling.

This chapter hopes to demonstrate the analytical value of shedding the
conventional view of the relation of discretion and law in favour of the alternative
approach taken in this thesis. Rather than considering discretion as the absence of
governance, discretion is considered here as a powerful form of governance, one which
facilitates the translation of certain social concerns into exclusionary immigration law and
policy and sovereign and coercive practices.

Discretion has always been a key component of the administration of Canadian
immigration law and policy, exempt to a startling degree from the eyes of the courts and
the public. The exercise of this discretion has been inflected by shifting, historically
specific discourses. In the past, discretionary powers facilitated the exclusion of large numbers of people on a variety of different discriminatory grounds that today are no longer judged to be socially or legally legitimate. As is argued in this thesis, in the present day, social concerns about criminality, inflected and constituted in varying degrees by racism and nationalism, provide the basis and justification for exclusionary immigration policies to a degree unprecedented in Canadian history.

This chapter begins by unsettling the conventional view of discretion and thereby sets the stage for the empirical and analytical exploration that follows.

II) The Conventional Legal View of Discretion and the Emergence of Alternative Approaches

When law ends, discretion begins, and the exercise of discretion may mean either beneficence or tyranny, either justice or injustice, either reasonableness or unreasonableness.

For the most part, contemporary legal theorizing on discretion continues to reflect the liberal preoccupation with the rule of law. The conventional view of discretion expresses the liberal legal conceit that discretion is the unruly shadow of law: a space which is to varying degrees unconstrained by legal or legalistic rules. In this space, essentially autonomous individuals 'freely' make decisions that ideally reflect rational, reasonable and objective calculations and aim to apply general legal rules to individual cases. A standard dictionary defines discretion as "the power or right to decide or act. according to one's own judgement or choice." In the context of the administration of public law and policy, the 'freedom to choose' of autonomous decision-makers is defined in natural opposition to the constraints imposed by legal rules. Indeed the same dictionary defines 'arbitrariness', the most prevalent concern associated with discretionary decision-making.

---


3 Random House 1987:563
making, as: "subject only to individual will or judgement, without restriction: contingent only on one's discretion: an arbitrary decision... having unlimited power: uncontrolled or unrestricted by law: despotic: tyrannical ...". The idea of discretion is thus heavily inflected with liberal assumptions and ideals relating to the power of law, autonomy and freedom of choice. Whether discretion is regarded benevolently or critically, its essential and inextricable binary relationship to law is largely taken for granted.

Ronald Dworkin theorized judicial discretion. As already noted, his 'doughnut analogy' expresses a view of discretion which rests upon the opposition of law and discretion. This view assumes, in its most extreme form, that law and discretion are discrete and distinct entities that are negatively correlated: more law means less discretion and less discretion means more law. Discretion is regarded and treated as a residual category, existing only in the absence of law. It also implies that discretion is itself a rather uninteresting space, a 'hole in the doughnut'. and that the only meaningful constraint on discretion is law.

In her work on discretion, Nicola Lacey challenges and seeks to unsettle the dominance of the legal paradigm which she defines as follows:

...the subjection of areas of human conduct and practice to regulation according to clear, prospective, publicly announced general rules or rule-like standards. Problems are typically seen as arising from ambiguities or 'gaps' in the rules, calling for clearer interpretations or further legislation

\---

4 Random House 1987:107

5

For fear of misrepresenting Dworkin's conception of discretion, it should be mentioned that in fact, Dworkin's conceptualization of discretion aimed to complicate and challenge this view by proposing that even where there is no explicit laws or rules which govern a decision, the constraining reach of the legal principles of the rule of law extends well beyond the written laws. Dworkin's account of discretion actually explains away discretion by arguing ultimately that discretion by its very nature is subject to its own limitations. Therefore he argues there is really no such thing as 'absolute' or 'unfettered' discretion. Judicial decision-making is always constrained by legal principles. Despite this positivistic departure, Dworkin's conception of discretion still rests upon its binary opposition to law; the only difference is the expansive way in which he defines 'law'.

or quasi-legal action. Disputes are seen as calling for resolution on the basis of given rules and according to standards of due process. This approach is closely associated with the ideal of the 'rule of law' and hence with liberalism as a doctrine of political morality.\(^6\)

In the current political debates on discretionary decision-making in the administration of detentions and removals, critics repeatedly voice their concerns about the prevalence of arbitrary and inconsistent decision-making occasioned by broad allocations of discretionary powers. Critics charge that arbitrary decision-making, that is decision-making that is not guided by law or rules, allows for the influence of racism, sexism, homophobia and other discriminatory and irrelevant considerations in the making of a decision. For many, the solution seems logically to lie in the creation and application of more rules. Indeed, this is the approach given strong expression in a major recent report discussed below. Arguably, current political thinking on these matters hasn't moved much past the 1960s. when discretion re-emerged as a 'problem'.

In the 1960s, discretion was increasingly regarded by legal scholars as a serious real and potential threat to individual justice. Discretion became increasingly understood as synonymous with tyranny and caprice - until that time, discretion had been generally regarded benevolently, as a 'humanizing' device to permit the general rules of law to be adapted to the separate circumstances of individual cases. One of the first major critiques of discretionary administrative decision-making was written by an American lawyer, Kenneth Culp Davis. His major and influential work *Discretionary Justice: A Preliminary Inquiry* \(^7\), published in 1969, issued out of his important insight that justice for individuals is administered more outside the courts than in them. Rather than as a clear cut binary, Davis conceptualised the relationship between discretion and law in terms of a scale or continuum - with absolute discretion at one end and absolute law on

---


the other. Davis was concerned with the potential for injustice associated with discretionary administrative decision-making. While he acknowledged the need for discretion in order to achieve individualised justice, he argued that the extent of discretion wielded by public officials was too vast. His proposals aimed to eliminate ‘unnecessary’ discretion and limit, constrain and structure ‘necessary’ discretion. He advocated the use of precise rules to ‘confine’ discretion thereby “eliminating and limiting discretionary power...fixing the boundaries and keeping discretion within them.”

Discretion is a problem for Davis only when it is inadequately regulated by rules.

In many respects, Davis’ work echoed though modified the arguments made by A.V. Dicey in 1915 regarding the choice facing democratic societies between ‘ordinary’ law and courts and ‘arbitrary’ and discretionary powers of public servants. Dicey associated discretion with arbitrary might and coercion: discretion in his view was the ‘antithesis of law’. It stands in direct opposition to the conventional view of the rule of law which dictates that people should be governed by laws not by arbitrary individual decisions.

Dicey’s was the dominant traditional view of discretion in the late 19th and early 20th century. With the rise of the welfare state, this view was replaced by the more benevolent view until the 1960s when Davis and many others began to return to these preoccupations. Discretion was then again viewed as the antimony of the rule of law. Discretion was associated with tyranny, arbitrariness and caprice whereas the rule of law assures certainty, fairness and consistency. As such, discretion has tended to be constructed as a problem to be acted upon; to be eradicated, constrained, limited or structured. From the conventional liberal legal perspective, discretion poses particular problems when it threatens individual rights and individual justice. Accordingly, as put by Lacey, “...the issue of control is often represented in terms of the protection of the

---

8 Davis. Discretionary Justice. 1969: 55

individual from state or bureaucratic abuse."\textsuperscript{10} While most acknowledge that its presence is necessary, indeed inevitable, given the complexity and diversity of the administration of the modern liberal state - nonetheless, its use (or abuse) has, since the 1960s, been a matter of growing concern.

Not only are laws and rules considered to be the only meaningful check on discretion but where there is a 'gap' in them, it is assumed that the individual decision-maker is afforded the opportunity to flex their essential autonomy and to be, to varying degrees, 'free to choose' between possible alternatives. Davis' definition of discretion is still widely cited, despite its generality: "A public official has discretion whenever the effective limits on his [sic] power leave him [sic] free to make a choice among possible courses of action or inaction."\textsuperscript{11} Generally speaking, most definitions of discretion rest upon the liberal assumption that the decision-making subject is fundamentally and inherently autonomous and that discretionary decision-making is thus a matter of (free) choice - even if it is a choice which is guided by external influences.

A succinct example of this traditional view is provided by H.C. Black, who asserted in 1968 that agents who exercise their discretion, do so "according to the dictates of their own judgement and conscience, uncontrolled by the judgement and conscience of others."\textsuperscript{12} As put by Bell:

\begin{quote}
(A)most any definition of discretion starts with the notion of choice ...
. The central aspect of this notion of choice is...the degree of self-
determination which the actor possesses as a consequence of the
responsibility which he [sic] has for achieving the success of a
particular enterprise. Tied closely with this aspect is the freedom the actor
\end{quote}

\textsuperscript{10}Ibid..367

\textsuperscript{11}Ibid..4

has from external constraints. 13

While the debates around the issue of discretion in the context of administrative decision-making reveal that these three assumptions (the primacy of law: law/discretion: and autonomy and choice) are still firmly underlie public and political debates and policy recommendations. recent scholarship on the issue of discretion reveals that these assumptions are beginning to be scrutinized and challenged. This study aspires to contribute to this more recent literature by developing and applying an historically sensitive. discursively oriented. social and political analysis of the exclusionary uses discretionary power in the development and enforcement of Canadian immigration law and policy.

In Davis' acceptance of the discretion/law dichotomy. Davis' solution to the 'problem' of discretion was to make administrative decision-making more rule bound and legalistic. assuming not only that rules and laws entail no element of discretion but also that the imposition of rules will necessarily result in increased individual justice. These assumptions reflect the more general and pervasive liberal legal conceit concerning the importance of the rule of law to the attainment of individual justice. It is assumed that procedural justice necessarily translates into substantive justice. Further. Davis' work continues the longstanding preoccupation of liberal legal thinkers with individual justice. D.J. Galligan. in his substantial study of legal discretion takes issue with a number of Davis' arguments. Central among these is his important observation that Davis excludes any consideration of policy-making in his definition of discretion. As put by Galligan. policy-making is "...the very heart of the discretionary process"14 This view is shared by Peter Manning who observes that policy. like legal rules. acts "...as one of the constraints

13 John Bell "Discretionary Decision-Making a Jurisprudential View" in Hawkins (ed.) The Uses of Discretion. 1992:93

in the context or field within which individual decisions have to be made."\(^{15}\)

The law/discretion dichotomy expressed in Dicey’s work and reasserted by Davis has also been increasingly subjected to critical scrutiny. Contemporary critical scholars now increasingly acknowledge that there is indeed no clear distinction between discretion and law. The law-discretion dichotomy has been increasingly unsettled though not totally undermined. Legal scholars concede that laws and their application are suffused with discretion and that administrative discretion is in fact extensively curbed by rules. As put by Madame Justice Beverly McLachlin: “Things are not as simple as Dicey perceived them. The law is not as certain as he would have it. nor are administrators as arbitrary.”\(^{16}\)

Lacey summarises the limitations of studies of discretion which derive from the legal paradigm. She argues that the legal scholarship on discretion tends to emphasise legal views and legal solutions over administrative ones. It is preoccupied with studying discretion in adjudicative settings and rests on a firm belief in the courts and adjudicative procedures. Emphasis tends to be placed on the importance of judicial review which in turn leads to an overemphasis on procedural justice at the expense of substantive justice. As Lacey points out, injustice can occur (as indeed can substantive justice) despite procedural controls. Ultimately, Lacey objects to the tradition of analytical jurisprudence, such as that engaged in by Dworkin, which tends to “...reinterpret the nature of the world by essentially imposing legal or quasi-legal categories of thought.”\(^{17}\) It is in light of these criticisms that Lacey calls for an interdisciplinary, pluralist approach to the study of discretion: an approach which views discretion in the context of legal, economic, political and social power.


\(^{17}\) Nicola Lacey “The Jurisprudence of Discretion: Escaping the Legal Paradigm” in Keith Hawkins (ed.) The Uses of Discretion. 1992: 294
Joel Handler is a legal scholar who, like Davis, is concerned about the potential for injustice which discretionary decision-making brings in tow. However, Handler does not share Davis' optimism regarding the effectiveness of legal or 'law-like' rules to govern discretion and thus maximize individual justice. Neither does Handler share Davis' view of the inherent threat of discretion. Rather, Handler analyses how "discretionary authority can be harnessed to advance social policy by serving the needs of powerless groups." 18 Handler is particularly attentive to the question of unequal power relations and unequal resources in the context of discretionary decision-making. He makes the obvious, but surprisingly rare, observation that solutions which call for more rule of law, more legality, more due process are of limited application "...especially since procedural due process protections are such a problematic remedy for the vast majority of dependent persons." 19 Handler argues that discretion is inevitable and that "...as long as large social agencies are serving and regulating the disadvantaged, then the problems of unfair power in the exercise of discretion must be addressed." 20

Discretion is ultimately a political issue, not simply a legal one. Lorne Sossin's work on discretion, administration and the welfare state shares Handler's concern with identifying ways of making the use of discretion in administrative contexts more responsive to the needs of the disadvantaged. Sossin, like Handler, is concerned to identify ways in which discretion can be employed to advance the interests of the disadvantaged. As put by Sossin:

Discretion represents an important and long undervalued means of integrating human relations into the administration of public authority. The question is not whether to allow discretion (because it is both endemic and indispensable to the functioning of the welfare state), but rather what kind of discretion should be fostered, and whose participation should be

18Keith Hawkins (ed.) The Uses of Discretion, 1992: 292

19Joel Handler "Discretion: Power, Quiescence and Trust" in Keith Hawkins (ed) The Uses of Discretion, 1992:360

20Ibid..360
Sossin is dismayed with the 'depoliticization' of the public sphere. He calls for the increased democratisation of administration by engaging the public in its processes. The problem with public administration, in Sossin's judgement, is that it is "closed off to [the public] as a conduit of political participation". Sossin finds Dworkin's 'hole in the doughnut' analogy unsatisfactory. In its place he offers up a 'sponge'. He states:

I find the emptiness in the centre of the doughnut an uncompelling representation for discretion's porous potential. Discretion is shaped by far more than the formal legal structures which surround and legitimate it. In a sponge, there are a variety of holes and substance in between, and the two are impossible to detangle. The sponge, however, can absorb and retain the fluids with which it comes into contact. This is how I believe the relationship between discretion and law ought to be viewed - not simply as a particular form of authority - but as a potential forum for politics.

Sossin advocates the pursuit of a 'communicative community' in the Habermasian sense. However, he, like Handler, is cognizant of the effects of unequal resources and positions of dependency on the attainment, or at least the approximation of the ideal of an inclusive, communicative public sphere. As he puts it, "While Habermas was surely right when he asserted that '[m]oney and power can neither buy nor compel solidarity and meaning', they certainly can and do skew the communicative potential of a contemporary public sphere." Sossin's attention to material inequalities is important. It should however be noted that there is in Sossin an enormous assumption that greater popular involvement equals greater justice.

The legal conceit expressed by Dworkin's doughnut analogy has resulted in most

---


22 Ibid. 9

23 Ibid. 12

24 Ibid. 13
attention being focused on the ‘surrounding belt of restriction’ rather than the ‘hole’ of discretion. Sossin argues that only by revealing the artificiality of the legal and political isolation of discretion can its ‘transformative’ potential be realised. In recent years there has been a growing interest in discretion-centred studies (as opposed to ‘law/rule-centred’). This is a positive development. Lacey argues that traditional legal approaches to law, and their tendency to treat discretion as a residual category, has meant that relatively little attention has been paid to how decisions are actually made and when attention has been paid to actual decision-making it has tended to focus on judicial discretion and judicial review.

Studies of ‘how decisions are actually made’ in other than judicial settings, have tended to come from the social sciences. Where legal scholarship has tended to privilege the law side of the dichotomy, social scientists have tended to focus on the discretion side privileging, in the process, social forces. In this, social science studies contribute to the gradual ‘de-centring’ of law by denying its essential primacy and by focussing analytical attention on ‘extra-legal’ or ‘non-legal’ influences on discretionary decision-making in particular contexts. The dividing line between law and discretion has been rendered less definite and the primacy of law less certain, however, the fundamental binary remains intact.

Social scientific approaches to discretion tend to emphasize that discretion is shaped by a myriad of so-called ‘extra-legal’ or ‘non-legal’ forces: that rules/law are but one of many possible constraints or influences on discretionary decision-making. While not really escaping the dichotomy, these studies do contribute to the gradual de-centring of law in the public, political and academic spheres. They seek to shift attention away from law and focus on the ‘extra-legal’ influences on discretionary decision-making. They acknowledge and take seriously the proposition that, contrary to the impression conveyed by Dworkin, legal rules are not the only important or effective constraints on discretion. As put by Bell.

(F)ar from being the uninteresting ‘hole’ in legal regulation, discretion is the centrepiece of the institutional edifice to which legal rules play a subservient role of setting the boundaries...the importance of legal controls
must be viewed in relation to other social controls which operate in a particular area."25

While social scientists might be faulted for underestimating the influence of law/rules in discretionary decision-making, the contributions which they have made regarding the actual uses of discretion are illuminating. Further, where legal scholars have attended to actual uses of discretion, they have tended to focus on adjudication and, as noted by Galanter, this focus on adjudication is "...vastly disproportionate to its prominence as a source of rules."26 Traditionally, social scientists had been more interested in law and justice outside the courts. Although, as observed by Hawkins, over the last 25 years sociologists, criminologists and political scientists have been increasingly interested in discretion in legal decision-making.27 For example, an important contribution was provided by D. McBarnett who observed that discretion is not only present in legal decision-making, but it is in actuality functional for the legal system in that it allows for the management of the lack of fit between legal rhetoric and reality.28 Lacey also observes that social scientists have shed light on other uses of discretion, for example: obscuring the lack of consensus or ambiguity about official policy; pre-empting the use of formal legal controls; or bestowing political and administrative power such as that wielded by immigration decision-makers.29

Some social scientists have approached discretionary decision making as a form

---


of administrative behaviour. One strain of analysis from this functionalist perspective is called the 'rational actor model' or 'classical rational model' of decision-making. This perspective assumes a fundamentally rational, free-thinking and autonomous conception of the individual decision-making agent. In this view, decisions are made in order to achieve the goals of an organization using the best available information. As described by Hawkins, in this view a decision "...is treated as a choice made by an individual based on unproblematic 'factors' or 'criteria' which may be deduced from the creation of a putative relationship between observed input (information) to a decision-maker and observed output." The usual criticisms of this approach are that: it assumes a congruence between the interests of the decision-maker and those of the organization; it assumes a clear and unproblematic and singular organizational goal or set of goals; and it assumes that decisions are made by autonomous individuals who are unconstrained by any other factors or influences. The full implications of this last criticism are rarely explored, namely that the fundamental premise of 'choice' and 'autonomy' belies an uncritical and taken-for-granted understanding of individual subjectivity; a criticism which, as we have seen, bedevils liberal legal analyses as well.

Additionally, the rational actor model thus fails to consider other 'non-organizational' factors which guide and shape discretionary decision-making. Influenced by this criticism, many social scientists have become increasingly interested in the importance of the broader context of decision-making and the various social, economic, political and legal factors which may guide (or shape or limit or determine - depending on your perspective) discretion. An expression of an extremely deterministic analyses of discretionary decision-making is provided by M. P. Baumgartner who asserts provocatively that discretion is a myth.31

Baumgartner proposes that the legal view that laws are the only meaningful

30 Keith Hawkins. "The Use of Legal Discretion: Perspectives from Law and Social Science" in Hawkins (ed.) The Uses of Discretion, 1992:22

constraint on discretion has obscured the role of social forces in constraining or shaping discretion. Baumgartner speaks of ‘social laws’ which she argues determine administrative discretionary decision-making in an extremely predictable and consistent fashion. She even asserts that social ‘laws’ result in more certainty that legal rules because social laws are general and transcend time and jurisdiction. In Baumgartner’s view, due to the functioning of social laws, discretion is synonymous with discrimination. Baumgartner, like Dworkin, presents an extreme analysis of discretion, however where Dworkin assumes the expansive and deterministic nature of legal laws, Baumgartner substitutes social laws.

There is now as well a substantial social science literature that has focused specifically on police discretion. Most of these studies are preoccupied with identifying the ‘factors’ that affect the uses made of the discretionary decision-making of the police and with measuring their relative statistical importance. Unlike the more ‘naturalistic’ approaches mentioned above, these studies do little in the way of critically examining the fundamental assumption of the autonomous decision-maker. Discretion is understood to mean ‘freedom to make choices’ or ‘power to choose’. For example, the policing literature on discretion subscribes for the most part to a ‘rule-based’ conception of and framework for, action\(^\text{32}\): just as we have seen with the legal analyses, the world of decision-making becomes divided into ‘legal’ and ‘extra-legal’ criteria which influence police decisions, in particular the decision whether or not to lay a charge. As summarised in a recent text on Canadian policing:

> Ultimately the police must have the discretion to decide whether or not to lay a charge. The question is which criteria the police will use when enforcing the law, not whether they will use discretion. Very simply, the criteria police use can be classified into four groups: criteria relating to the offence, criteria relating to the offender, criteria relating to the officer, and criteria relating to the setting in which the offence was committed. A few of these criteria, such as the seriousness of the offence and the availability of evidence are appropriate criteria as demonstrated by the fact that they

\(^{32}\)The term ‘rule-based conception of action’ is borrowed from C. Shearing and R. Ericson “Culture as Figurative Action” British Journal of Sociology. Vol.42. No.4. 1991:481
are authorised by law: however, many other criteria police use, such as the sex, race and sexual orientation of the offender, are prohibited by law and therefore should not be used by police. 33

Clearly, the rule-based conception of decision-making is pervasive in the policing literature and with it, the uncritical acceptance of the very notion of ‘freedom to make choices’ and the underlying assumption about the inherent autonomy of individual subjectivity which such notions presuppose. Also endemic is the zero sum concept of rules and discretion: more rules means less discretion and vice versa. The following observation made in a recent text on policing is typical: “The fewer the rules about handling incidents and situations, the more discretion officers have” 34 Furthermore, most of the research in this area is quantitative, preoccupied with identifying and measuring the relative statistical importance of different ‘legal’ and ‘extra-legal’ criteria on police officers’ discretionary decision-making. The inherently autonomous decision-making agent is taken for granted.

A provocative departure from the conventional preoccupations of this particular field of study is provided by Shearing and Ericson. They, like Manning, make use of ethnomethodological insights and semiotics in their analysis of police action and culture. Shearing and Ericson bring to the discussion an attention to the construction of subjectivities and the role of cultural narratives in this construction thereby challenging what they term the ‘hegemony of the rule-based paradigm’. As put by the authors:

In studying policing, sociologists have examined the fit between legal rules, viewed as instructions, and police decisions. Typically they report that police officers deviate from these legal instructions...Instead of using these findings to question the rule-based paradigm, however, they have accepted as axiomatic the belief that all action is rule generated and concluded that there must be some other set of rules that is generating


police action.\textsuperscript{35}

This observation is generally applicable to a large number of sociological studies of administrative decision-making and indeed of sociological studies in general. Attempts to identify 'law-like' or 'rule-based' explanations of decision-making are widespread. Baumgartner's study discussed above provides a typical, albeit extreme, example of this. Shearing and Ericson attempt to provide a theoretically more sophisticated analysis of police agency - one that does not simply explain action by reference to individual choice or (assumed) rules. In so doing, they bring the difficult questions of the constitution of subjectivities and the artful construction of realities to the fore. In this, their approach is similar to Manning's analyses of administrative decision-making.\textsuperscript{36}

Manning approaches the uses of discretion from a 'naturalistic' perspective. The naturalist approach attends to the importance of context in understanding the nature of discretion. It is concerned to study how decisions are actually made: the basis of these decisions and their meaning and significance to the agents who undertake them. However, in this approach, decisions are not considered to be the product of individual choice or rational, autonomous judgement and planning. Rather, these social scientists have focused on the importance of structures of meaning ('frames') and contexts of decision-making ('horizons'), including organizational, social, political and economic.

Manning, like Shearing and Ericson, is similarly opposed to rationalistic and individualistic views of discretion. Emerson and Paley's work on administrative decision-making\textsuperscript{37} also shares this orientation. They observe that a case cannot be analysed independently of context: 'facts' and 'files' are 'artfully' assembled and this involves

\textsuperscript{35}C. Shearing and R. Ericson, "Culture as Figurative Action" \textit{British Journal of Sociology}. Vol.42, No.4. December 1991:483

\textsuperscript{36}P. Manning, " 'Big Bang' Decisions: Notes on a Naturalistic Approach" in Hawkins (ed.) \textit{The Uses of Discretion}. 1992

\textsuperscript{37}R. Emerson and B. Paley "Organizational Horizons and Complaint Filing" in Hawkins (ed) \textit{The Uses of Discretion} 1992
"...not only the construction, but the selective reconstruction of reality." The 'naturalist' approach is strongly influenced by the insights of ethnomethodology and semiotics. It begins to explore how decisions are 'constructed' as opposed to 'made' by decision-making agents. It complicates the generally taken-for-granted understanding of subjectivity and rationality. Moreover, it does not seek to identify new 'universals' to explain decision-making but rather sees decision-making as an activity constructed by a myriad of shifting and dynamic influences - influences which do not merely shape decisions but which actually constitute them.

III) Administration and Judicial Review

In any policy discussion of discretionary administrative decision-making, the question of judicial review is frequently raised. Again, this preoccupation with judicial review is a further indication of the continuing dominance of the law/discretion opposition and its prevalence in public and political discourse. This is particularly relevant to the issue of detention under the Canadian Immigration Act which provides the primary focus and example for this discussion of discretion. The debates around the question of immigration detention always include consideration of the extent of judicial review available as a potential check on the abuses of discretionary decision-making in this area.

With respect to the question of judicial review, it has been observed that judicial self restraint is evident in all issues respecting immigration, national security and/or defence. The most extreme historical examples of this kind of restraint include the deference by the courts to the broad tracts of administrative discretion used by the state during wartime: measures which were extreme, coercive and discriminatory. This judicial deference subsided somewhat after the war. However, it has been observed that the courts continue to be reluctant to intervene in administrative decision-making when the decisions rendered are related to questions of national security and public order: as

38Ibid., 124
put by Cowan. "...considerations of national security and public order still influence judicial review." Cowan goes on to note that while the courts maintain a 'residual' jurisdiction to ensure procedural fairness, they have held that "the resolution of the conflict between national security and individual freedom was a non-justiciable issue. one for the Minister. who was responsible to Parliament alone. to decide.""40

An alternative approach to thinking about the relations between law, administration and discretion is developed by Peter Fitzpatrick. In contrast to treating the limited availability of the remedy of judicial review for immigration decision-making specifically, and indeed of administrative decision-making more generally, as a residual problem stemming from the inherent opposition and conflict between law and administration. Fitzpatrick examines this tension as a productive and positive one: one that results in the 'mythical mutuality' of law and administration.

In his work entitled The Mythology of Modern Law41. Peter Fitzpatrick provides an account of the relation between law and administration that is particularly compelling. Fitzpatrick develops intriguing and persuasive analyses of the historical development of the relations between law and administration and the apparent binary which they represent. Fitzpatrick argues that instead of an antinomy existing between law and administration, there is "...an operative compatibility between them".42 Fitzpatrick is responding here to the claims made by scholars, in particular by Unger, of 'the death of law'. Fitzpatrick observes that the standard story is that the rise of bureaucracies and expanded administration has led to increased inconsistency. increased discretion and reduced generality: in this tale "...[t]he rule of law - as general in application. predictable

---


40 Ibid. 32


42 Ibid. 149
and autonomous - is thus undermined.\textsuperscript{43}

Fitzpatrick approaches the law as myth. Approached in this way, law and administration may be regarded as having a 'mythic mutuality'. Fitzpatrick's argument is that administration and law are mutually, mythically, coherent. That is, the limits of each are supplemented by the reach of the other and, in the process, the mythical existence of each is reaffirmed. In his words:

... [the] pervasive and tentacular penetration [of administration] inevitably displaces or at least marginalizes law. Yet this...is only part of the story. If we consider the nature of modern administration itself, we find that it is bound by certain operative limits. It is in those very limits that the rule of law assumes a dynamic existence, one distinct from and opposing an administration to which it forms a necessary supplement. Moreover, in this relation to administration, law itself is limited by and dependent on administration. A contradiction involved in which law and administration are integral and yet necessarily opposed. Law and administration in their mythic mutuality limit each other yet sustain the claim of each to be unlimited. \textsuperscript{44}

Fitzpatrick is taking issue with the widely held view that the demise of law originated in the late nineteenth century with the growth of extensive state administration. He rejects the view that administration has undermined law from the outside so to speak. arguing instead that the allegedly 'corrosive' effects of administration were in fact 'integral' to modern law.

Fitzpatrick explores Foucault's proposition that modernity entailed a shift from sovereign legal power to disciplinary rule. Fitzpatrick argues that the conventional tale which holds that the rule of law, in all its certainty, predictability and generality, was 'undermined' by an ever expanding administration characterised by discretion, particularity and variance, rather than evidencing the death or imminent demise of law, actually speaks to its continued power and influence. His argument is that the rule of law derives its very existence in contradistinction to the 'particularistic and pervasive powers' of administration.

\textsuperscript{43}Ibid..147

\textsuperscript{44}Ibid..149
of administration. Thus, it is not simply that law has been displaced to the margins or the periphery "...confined to some largely symbolic or procedural oversight of the operation of administration" but in fact, it is precisely because of these particularistic and pervasive powers of administration "...that the rule of law can be maintained in its aspects of universality and equality and seen as marking out fields for free action." Accordingly, in Fitzpatrick's analysis, if law were substantially to counteract administration, it would in fact be undermining the conditions of its own existence: "administration is the necessary 'dark side' of law".

Fitzpatrick supports this proposition by examining judicial review of administrative action. He aims to show that despite the fact that judicial review is held to be "...operatively the ultimate expression of the rule of law and its dominance over administration" it actually has "heightened the limits of law". He provides evidence which speaks to both the inherent limitations of the rule of law vis a vis administration and the converse. He outlines many of the mechanisms and requirements that preclude law's interference with administration, including for example the requirement that judicial review limit itself to questions of procedural fairness or 'improper' uses of discretion and the vague, discretionary and and watered down application of the requirements of 'natural justice'. He also describes those checks that are present in the field of administration which prevent or 'obviate: recourse to the law. Fitzpatrick

45Ibid..154

46Ibid..154

47Ibid..154

48

My reading of the legal literature on the judicial review of administrative decision-making is consistent with Fitzpatrick's thesis, more specifically, the law places considerable limits on the availability of judicial review for administrative decision-making. The limited availability of judicial review is particularly pronounced in the context of immigration and parole decision-making. While this will be addressed in a general way a little later on in this chapter, for a much more comprehensive discussion of administrative law, discretion and judicial review see Charter of Rights and Administrative Law, Law Society of Upper Canada, 1986.
concludes that the relation between law and administration is one of interdependence and mutual affirmation. Fitzpatrick argues that "...law, in short, provides a guarantee that everything is really so." He explains:

The scope remaining for law consists only in dealing with aberrations. Law is thence responsibly self-limiting. As an institution, it constitutes itself and its interventions as occasional and discontinuous. Law is not, however, simply responsive to the nature of things. It evokes, affirms, even creates the normal and authority as normal. Through law’s shaping and dealing with the exceptional and the aberrant, with what is outside the properly administered world, administration is rendered normal and right.50

In this way, Fitzpatrick demonstrates the dynamic processes by which we become ‘reassured about bureaucratic organization’ .51 This view can be read as an elaboration of G.E. Frug’s more conventional critical analysis of public administration. Frug was concerned to ‘undermine the ideology of’, and ‘expose the false consciousness’ induced by bureaucracy. Writing in 1984, Frug also made the shrewd observation that the scholarly, legal and political preoccupation with the judicial review of bureaucratic decision-making is precisely what “defines, perpetuates, explains, justifies and reassures us about bureaucratic organization.”52

50 Ibid., 160
51 Ibid., 160
In his provocative work entitled "Welfare, Rights and Discretion" Robert Goodin makes a persuasive case for abandoning the longstanding preoccupation with the problems of discretion and legal avenues of addressing them. Goodin argues that the logical opposite of enjoying discretion is being bound by a rule. He observes that this binary leads to the unjustified conclusion that the problems of discretion can be resolved by the application of a rule. In his words:

...(L)ogically, the natural response to finding that certain problems are inherent in discretion is to impose rules in the place of those discretions. The assumption that that will automatically solve the problem of discretion, however, entails the unwarranted presumption that every problem necessarily has a solution. Saying that the problems ...are inherent in discretion means that those problems can be resolved, if at all, only by curtailing that discretion through rules. But it is perfectly possible that those problems cannot be solved in that way either.

Goodin argues that ultimately the problems associated with discretion as they have been conventionally understood (arbitrariness, inconsistency and discrimination) are largely 'insurmountable' and 'ineliminable'. He makes the important observation that these problems conventionally associated with discretion may also be associated with rules. While Goodin's analysis of discretion does not ultimately disrupt the law/discretion dichotomy, he does argue that the logical assumption that issues from this dichotomy—that the problems associated with discretion can be remedied through the application of rules - is misguided. Goodin argues that the only problem associated with discretion that can be addressed through rules is the problem of manipulation or exploitation by decision-makers. Apart from this possible remedy. Goodin points out that all the other problems commonly associated with discretion are also in evidence with rules and laws. He states that rules, "...cannot (necessarily, or without substantial costs in other respects) prevent intrusiveness, arbitrariness or insecurity. For much the same reasons that

---


54 Ibid., 250
discretionary decisions must display those attributes. rule-based decisions can, and probably will.  

Goodin’s work is an important contribution to the study of discretion and is particularly germane to the analysis developed in this thesis. He meticulously makes his case against rule-based problems to discretion through a persuasive critique of the recourse to law that the law/discretion dichotomy implies. Goodin’s work is important for another reason as well. After making his case for circumventing the problem of discretion and the privileging of rules in the resolution of such problems, Goodin draws attention to the link between discretion and the particular purposes which underlie both rules and discretion in a given context. Goodin’s empirical site for his discussion of discretion is the provision of welfare assistance. He argues that the problems which appear to arise in relation to discretion, actually arise because of the underlying purpose of welfare decision-making to distinguish between needy and deserving claimants and the needy and undeserving. It is this purpose, rather than discretion per se, that underlies the problems associated with discretionary decision-making. Goodin concludes by proposing that rather than dispensing with, or seeking to dispense with discretion, it is this purpose that should be dispensed with. More specifically, Goodin proposes that:

Imagine, in contrast, a world in which officials were guided only by the first half of that twin obsession: suppose officials are anxious to ensure that everyone who needs/deserves benefits gets them, but that they are utterly unconcerned to ensure that only they receive them.  

For the purposes of this study. Goodin’s work is valuable not only because it whittles away at the law/discretion dichotomy, but perhaps even more importantly because it makes the link between discretion and discretionary context: the underlying purpose of decision-making. In the context of immigration detention decision-making, the problems associated with discretion are similarly derived from the underlying practical

55 Ibid., 259

56 Ibid., 259
purpose of the decision-making in question: to distinguish between desirability and undesirability in the context of prospective immigrants and between deservedness and undeservedness in the context of refugee claimants. Discretionary power in the context of immigration decision-making is, by definition, discriminating and preferential: what is of particular importance then, is the bases for and the effects of these preferences.

This study of discretion and immigration detention decision-making takes seriously the specific context of discretionary decision-making. It assumes that discretionary power cannot be studied in isolation or in the abstract. The nature, uses and effects of discretionary power must be considered in relation to the political rationales which govern and justify it, the specific institutional and organizational dimensions of its existence, the local and dispersed nature of its workings and the complex, overlapping and dynamic processes which it traverses.

The limitations of the discussion and debates around discretion, administration and law suggest the need for the development of new ways of thinking about discretion. This study of discretion in immigration detention-making attempts to contribute to this development. It seeks to avoid the problems that issue from the core assumptions which underlie the contemporary debates, namely the primacy of law, the law/discretion dichotomy and the notion of autonomy and ‘free choice’.

Understanding how exclusionary immigration decision-making takes place requires a different way of thinking about discretion. Rather than thinking of it as a relatively uncomplicated expression of individual agency unchecked, to varying degrees, by legal constraints, discretion can be thought of as a particular form of power characteristic of liberal forms of governance. The conventional view of discretion reproduces and reinforces the liberal idea of the autonomous individual agent: it serves to maintain and reproduce the distinction between law and administration; and it insulates many areas of governmental decision-making from serious critical scrutiny.

In this work, discretionary immigration detention decision-making is regarded not only as a negative and coercive power, but also and necessarily as a positive and productive power. Discretionary power is positive in the sense that it may be employed.
albeit in varying degrees, in the advancement of different objectives. Discretionary power is productive in the sense that it produces subjectivities and identities which are then regulated accordingly. Discretionary power is local and dispersed, both within governmental and non-governmental spheres. While it is true that those discretionary powers that are associated most closely to the state are likely more powerful determinative of outcome, discretionary powers are also noteworthy at non-state levels. For example, while the discretionary powers of legal aid to grant or deny funding for legal counsel for a detention review may not be equal to those of adjudicators who determine the outcome of individual detention cases, they are none the less still extremely important and operate in conjunction with other discretionary powers in the resolution of individual cases. Discretionary powers are thus local and dispersed in the sense that they exist both inside and outside the state and are dispersed along different networks, processes and agencies. Additionally, discretionary powers are local and dispersed in the sense that even within the sphere of government, different agents and agencies all engage in discretionary decision-making that is not, and arguably could not possibly be, comprehensively regulated by legal rules or departmental policies. Understanding discretion in this way allows for the exploration of the complex and dynamic relations and processes that together all impact upon the administration of state law and policy in the context of immigration detention.

In this view, discretionary power produces the figure of the autonomous, free-thinking decision-maker, in this context, the immigration officer or the adjudicator. Discretionary power is part of, and indeed constitutes their subjectivities. In the context of Canadian immigration detention and deportation policy, this productive discretionary power is then employed in repressive sovereign ways against "undesirable" (criminal, infected, destitute, uncooperative) and/or undeserving (for example the "bogus refugee") non-citizens. However, these negative and coercive uses of official discretionary power are also productive. The application of coercive and repressive sovereign sanctions against undesirable and/or undeserving non-citizens (detention, use of body restraints, deportation) serves to reproduce the power and "rightness" of sovereign penalty. Further.
the coercive and negative sovereign sanctions applied against non-citizens are also productive of a desirable and deserving citizenry (law-abiding, healthy, acquisitive and cooperative).

The discretionary power which governs both desirable and undesirable non-citizens in coercive and productive ways is inflected by dominant, historically specific discourses. In the present day, the dominant discourses which inflect the uses of discretionary power in the administration of the exclusionary provisions of the Act, in particular those provisions which sanction the detention and deportation of non-citizens, are the discourses of danger/risk and criminality. Certainly alternative discourses complicate decision-making in this area, most influential of these issue out of a human rights perspective and out of the liberal legal paradigm.

This proposition will be explored and developed throughout this thesis. For now, suffice it to say that in the present day undesirability is increasingly associated with criminality and danger/risk such that the detention and deportation of non-citizens is primarily justified on those grounds. Today, the archetypal undesirable citizen is the ‘criminal immigrant’ who represents a risk to the ‘safety and good order’ of Canadian society. The coercive regulation of the ‘criminal immigrant’ also serves to constitute and govern all non-citizens, not only those subject to its applications. More specifically, the negative and coercive sovereign powers of detention and deportation also represent the ever present looming spectre which contributes to the constitution and regulation of the figure of the law-abiding desirable immigrant. In this way it can be seen that negative and coercive uses of sovereign power, detention and deportation, interacts and coexists with less visible forms of regulation. It also can be seen how the particular use of discretionary power cannot be divorced from its specific context and the dominant discourses which inflect its uses.57

57 Discourses do not simply refer to ways of speaking or thinking about an issue, but also to ways of acting upon an issue; laws and policies, institutional and organizational arrangements, networks, strategies and processes.
IV) The Law and Policy Governing the Detention of Non-Citizens Under the

*Immigration Act*

Discretion is today and has always been a central feature of the Canadian Immigration Act. Discretionary decision-making, political and bureaucratic, is integral to the political and legal determination of desirable and undesirable immigrants and deserving and undeserving refugees. The conventional legal definition of discretion is "the power to choose between two or more courses of action each of which is thought of as permissible." 58 Administrative law has long been characterised by the degree of discretion afforded to its administrators. This discretion has been justified on primarily two grounds: a) discretion allows for the tailored application of general laws to individual cases facilitating the attainment of individualised justice and b) discretionary powers of decision-making are essential to the efficient and effective administration of legislation that affects huge numbers of people in complex and varying ways. Canadian Immigration legislation is suggestive, allowing for considerable scope for discretionary decision-making at each stage of the process. While the discretionary powers of criminal justice officials have been widely studied, monitored, debated and critiqued: not so the discretionary powers of Immigration officials. 59


59 A review of the scholarly literature on immigration detention in Canada is not included here largely because it doesn’t exist. A recent American exception to this general lacuna is provided by Janet A. Gilboy (1991, 1996). Despite the proliferation of scholarly works, both historical and contemporary, on the subject of Canadian immigration policy and practises, very few engage in any analysis of discretion and/or detention as objects of study. If mention is made of detention, it is usually limited to a brief description of the law and policy which governs its uses. Where discretion is regarded as a ‘problem’, the focus tends to be on the need to tighten up judicial review and due process. This lacuna is not insignificant for my thesis. The absence of such studies indicates that for the most part, scholars too continue to be preoccupied with legal remedies and have not moved beyond the limits of the law/discretion dichotomy and the liberal legal paradigm. Moreover, the absence of any theoretical or empirical research on Canadian immigration detention certainly represents a ‘gap’ which this thesis addresses.
The debates around the uses of discretion in the administration of governmental law and policy have understandably been particularly pointed when the decisions that are taken on individual cases entail serious consequences for the life and liberty of the person in question. The administration of the legislative provisions of the *Canadian Immigration Act* (forthwith referred to as the *Act*), and in particular those provisions dealing with the detention and removal of non-citizens entails just such dire consequences. The *Act* sanctions the use of coercive sovereign power against non-citizens who, as a result of a series of broadly discretionary decisions and a complex web of interactions between governmental and non-governmental discretionary powers, have been determined to require detention.

Immigration officers (who are civil servants employed by CIC) and adjudicators (who are order in council appointees to the independent adjudicative tribunal, the Immigration and Refugee Board) have broad statutory powers to detain non-citizens. In the field of immigration, parliament has established two main grounds that justify detention: 1) the person is likely to pose a danger to the public; and 2) the person is not likely to appear for an examination, an inquiry or removal. (Sections 80.1 and 103 of the *Act*). Immigration officers make the initial decision to arrest and/or detain. All detention cases must then be reviewed by a Senior Immigration Officer (SIO) within 48 hours of the initial detention. The SIO reviews the case to determine whether release is appropriate. In exercising their power, the SIO has considerable discretion in imposing terms or conditions of release on a person who has been detained. If the person is not released within 48 hours of their initial detention, seven days later, the decision to detain

---

60 Although 'fail to appear' and 'danger to the public' are the most common grounds for detention, other grounds exist under S.103.1. According to both the IRB *Guidelines on Detention* and the draft copy of the yet to be released CIC *Guidelines on Detention*, these grounds are however rarely applied. They relate to port of entry cases where a) a person is unable to satisfy an immigration officer with respect to that person's identity, or b) in the opinion of the Deputy Minister or a person designated by the Deputy Minister, there is reason to suspect that the person may be a member of an inadmissible class described in paragraph 19 (1) (e), (f), (g), (j), (k), or (l).
is again reviewed. however from this time onwards. the review is carried out by an independent immigration adjudicator who is a member of the Immigration and Refugee Board. a quasi-judicial independent tribunal. Adjudicators have the power to order the detention or continued detention of a person. They may also order that a person be released subject to terms and conditions which the adjudicator deems appropriate. including for example the payment of a security deposit or the posting of a performance bond. If. at the 7 day interval. the detention is continued. subsequent reviews take place at 30 day intervals until the case is resolved.

The courts ruled in Sahin (1994) that in addition to making a determination as to whether one or both of the two acceptable grounds for detention exist, the decision-maker, either the immigration officer or the adjudicator. must also consider whether continued detention accords with the principles of fundamental justice under section 7 of the Charter. This judgment should perhaps be quoted at length:

Although the power which is given to the Secretary of State in para.2 to detain individuals is not subject to any express limitation of time. I am quite satisfied that it is subject to limitations. First of all. it can only authorize detention if the individual is being detained in one case pending the making of a deportation order and, in the other case, pending his removal. It cannot be used for any other purpose. Second, as the power is given in order to enable the machinery of deportation to be carried out. I regard the power of detention as being impliedly limited to a period which is reasonably necessary for that purpose. The period which is reasonable will depend on the circumstances of the particular case. What is more. if there is a situation where it is apparent to the Secretary of State that he is not going to be able to operate the machinery provided in the Act for removing persons who are intended to be deported within a reasonable period. it seems to me that it would be wrong for the Secretary of State to seek to exercise his power of detention.

This means that while the Act does not limit the total length of detention, there are implicit restrictions. namely that the length of detention must be "reasonable". As stated


in the IRB Guidelines:

...if a detention appears unduly lengthy, the reasonableness of the delay should be considered in order to ensure that the detention is not in fact an 'indefinite detention'. Such detentions constitute deprivations of liberty that come into conflict with the principles of fundamental justice.  

In addition to these powers of detention, immigration officers also have broad powers relating to arrest without warrant. It has been frequently argued that these coercive powers of immigration officers actually exceed those of police officers. For example, immigration officers may detain someone if they are of the opinion that the person in question is not likely to appear for future immigration related proceedings. As put by Rivka Augenfeld, President of Table de Concertation de Montreal pour les Refugies during the 1997-98 Standing Committee hearings on detention:

...Immigration Officers have powers beyond the range of any policeman(sic) or soldier in this country. There is no policeman who can arrest somebody because he thinks the guy won't show up for a hearing next week. If the person doesn't show up, you issue an arrest warrant, but you can't decide Joe Blow isn't going to come next week. then go to his or her house, arrest them in their pyjamas and take them to detention.  

Despite these broad powers of arrest and detention, the decision to detain under the Act has not elicited anywhere near the same degree of critical public, political or academic attention as has the detention of an individual charged and/or convicted with a criminal offence. Both forms of detention represent extreme examples of the continuing 'coercive and 'bodily' powers of the sovereign. In a liberal regime, the deprivation of an individual's liberty by the state is considered to be a very serious matter and in the context of criminal justice administration the state takes great pains to check this use coercive state power against individuals. The criminally accused is protected by a myriad

---

63 IRB Guidelines. March 1998:3 (emphasis in the original text)

64 Standing Committee Hearings on Detention. Wednesday March 18, 1998:25
of legal safeguards and protections: the principles of fundamental justice, the rule of law, due process. State sanctioned deprivation of liberty is generally thought of in relation to the administration of criminal justice where the potential loss of liberty is justified on punitive grounds.

In contrast, according to official law and policy, the detention of non-citizens under the Immigration Act is preventative not punitive. This legal differentiation between preventative administrative detention and punitive detention is not likely to be readily apparent to those who are detained. On a general and somewhat abstract level, the designation of immigration detention as preventative means that it must be regarded as an exceptional measure. This principle emerges from statute and case law, and is encoded in, for example, the Canadian Charter of Rights and Freedoms, and the International Covenant on Civil and Political Rights.

This discursive distinction between punitive and preventative detention moves immigration detention from the ambit of criminal law into that of civil law and as such the merely 'quasi-judicial' nature of detention decision-making does not carry with it the same obligation to provide and apply the legal protections and safeguards which are required under criminal law. For those who are detained, the material consequence of this distinction is that while they may experience their detention as punishment, they are not accorded the degree of rights and protections as they would were they actually facing punishment. For example, legal rules of evidence do not apply, access to legal counsel is limited (there is no duty counsel available and until the late 1990s legal aid did not issue certificates for detention reviews), and the onus of proof is a reverse onus, that is to say it is the detained person who must bring forward evidence to prove that she should be released. The longstanding official legal distinction between punitive and non-punitive forms of state-sanctioned detention which underpins the reluctance of the courts to interfere with detention decision-making is an ideological distinction which contributes in a central way to the general lack of critical academic and political attention paid to

---

65 This is particularly likely to be true for those people who are being detained in correctional facilities along side convicted criminals who are being punished.
immigration detention.

The legal status of immigration administrative decision-making and the longstanding tradition of judicial restraint in this area of public law and policy is centrally supported and reproduced by the ideological concept of state sovereignty. The broad powers to arrest and detain under the Immigration Act are officially related to national social, economic and political aims and objectives. The historical legacy of state sovereignty and the conceptions of sovereign rights and duties which derive from this legacy continue to underpin and justify even these most coercive of the activities of the Department. The evidence is overwhelming, from Canada's early treatment of Chinese labourers, to the detention of the Japanese and Canadians of Japanese origin, that the people and the Canadian government have easily and comfortably accepted that the rights of non-citizens can be abridged with vastly more ease than those of citizens, evidencing the continuing hegemony of liberal conceptions of sovereignty and national security which justify the uses of coercive state power against non-citizens.

The coercive powers of the state to detain and deport non-citizens are largely justified by reference to the sovereign right control national borders and populations and the sovereign duty to protect its citizens from danger, to maintain public order and economic well-being. The objectives which guide the Immigration Act are laid out in s.3 of the Act. The subsections that are of particular relevance here are the following:

3. It is hereby declared that Canadian immigration policy and the rules and regulations made under this Act shall be designed and administered in such a manner as to promote the domestic and international interests of Canada recognizing the need....

f) to ensure that any person who seeks admission to Canada on either a permanent or temporary basis is subject to standards of admission that do not discriminate in a manner consistent with the Canadian Charter of Rights and Freedoms:

g) to fulfill Canada's international legal obligations with respect to refugees and to uphold its humanitarian tradition with respect to the displaced and the persecuted:

h) to foster the development of a strong and viable economy and the prosperity of all regions;

i) to maintain and protect the health, safety and good order of Canadian society; and
j) to promote international order by denying the use of Canadian territory to persons who are likely to engage in criminal activity.

The tension between, on the one hand, ensuring that the requirements imposed by the *Canadian Charter of Rights and Freedoms* and the obligations regarding human rights derived from international legal obligations are fulfilled and, on the other hand, the sovereign obligation to protecting the 'health, safety and good order of Canadian society' pervades the administration of the Act. Those who are vested with the responsibility of enforcing the Act tend to regard the obligations relating to the legal and human rights of non-citizens as a constant frustration in their ability to do their job. This is reflected in the official designation of these legal rights as 'impediments to removal'\(^{66}\). The detention of non-citizens under the Act is justified as an administrative prerogative of the state issuing out of its right to control its national borders and the nature of its population. As put by political philosopher Joseph Carens in making his persuasive case for open borders:

> Borders have guards and the guards have guns. This is an obvious fact of political life but one that is easily hidden from view - at least from those of us who are citizens of affluent Western democracies...most of those trying to get in are...ordinary, peaceful people, seeking only the opportunity to build decent, secure lives for themselves and their families...What gives anybody the right to point guns at them?

To most people the answer to this question will seem obvious. The power to admit or exclude aliens is inherent in sovereignty and essential for any political community. Every state has the legal and moral right to exercise that power in pursuit of its own national interest, even if that means denying entry to peaceful needy foreigners. States may choose to be generous in admitting immigrants, but they are under no obligation to do so.\(^{67}\)

The underlying principle of national sovereignty and its association with questions of national interest and the state's obligation to protect national citizens is a

---

\(^{66}\) IRB Guidelines on Detention 1998:2

powerful one, so powerful that it is often seem to ‘trump’ the human rights related legal principles which are also included in s.3 of the Act.

V) Judicial Review of Immigration Decision-Making

Any study of discretionary powers in the context of the administration of public law and policy must consider the relationship of administrative decision-making with the rule of law and the principles of fundamental justice as manifested in the principles and processes which govern the judicial review of administrative decision-making. It is a generally accepted view that judicial review provides an important check on discretionary administrative decision-making powers. The longstanding association of discretion with arbitrary decision-making has been conventionally linked in dichotomous opposition with the rule of law. That is to say, it is generally accepted that judicial review functions as a necessary top-down check on the potential for injustices in administrative decision-making. Within this conventional approach, judicial review operates in accordance with universal principles of fundamental justice whereas administration depends upon the legislated granting of discretion for its effective management.

In part, the lack of critical attention given to discretionary decision-making in the field of immigration can be attributed to the fact that the administration of the immigration act, including those provisions which sanction the use of coercive measures such as detention, is governed by Administrative law. Administrative law and policy has always enjoyed a significant degree of protection from judicial oversight. This stems in part from the historical understanding that administration relies for its effectiveness on the highly specialized expertise of its administrators, expertise which is not possessed by the judiciary. Judicial non-interference is particularly pronounced in the area of immigration decision-making. As observed by legal scholar, P. Bryden:

Immigration law, like the law surrounding parole and prison discipline, has had a reputation among people interested in administrative law as a sort of waste land in which judges have been loathe to apply the legal principles we normally associate with a sense of justice in Canadian public
Administrative law experts confirm this impression. J.G. Cowan, for example, observed:

Cases involving immigrants, prisoners, and parolees are examples of other areas of discretionary power in which the courts are reluctant to interfere. Statutory powers exercised by immigration and prison officials are quite wide and considerable scope has been given to the purposes for which they may be exercised and the factors that may be considered in reaching a decision, all of which are reflected in decisions in which the principles of natural justice have been modified significantly or ignored.

While the judiciary maintains a residual jurisdiction to review the procedural fairness and propriety to ensure that administrative decision-making operates within the general parameters of due process (that it is not arbitrary or capricious) and a rather watered-down version of natural law and the principles of fundamental justice, the substance of administrative immigration decisions is rarely interfered with by the courts. Under administrative law, discretionary decisions which are deemed to be 'purely administrative' in nature are not subject to judicial review. However, broad areas of decision-making under the Act are explicitly 'quasi-judicial' and as such are subject to judicial oversight. In practice, however, the availability of the remedy of judicial review is quite limited: the courts have long been loathe to interfere with the discretionary decisions made under the Act. In the 1997-98 Standing Committee hearings on detention, Paul Thibault, Executive Director, Adjudication Division, Immigration and Refugee Board (IRB), offered the following numbers: 16,000 decisions are made by adjudicators on an annual basis. The Federal Court sets aside about .05% of these. This works out to

---


about 8 cases annually which are set aside by the Federal Court.\textsuperscript{70}

\textbf{VI) Departmental Policy Checks on Discretionary Powers}

In addition to the checks on administrative discretion, albeit limited, provided by the courts, the uses of discretionary decision-making may also be constrained and shaped through the development and application of "non-binding" departmental policies and guidelines such as those recently issued by the Chairperson of Adjudication on the subject of detention. Departmental guidelines such as these are not binding, for that would represent an undue interference with the independence and discretion of the decision-makers. Rather, they articulate a "recommended approach" with the aim of "promoting consistency, coherence and fairness" in the making of decisions. However, even non-binding departmental guidelines elicit significant controversy and debate, revealing again the extent to which discretion continues to be a divisive and difficult issue in the context of the administration of the Act.

For example, in the fall of 1996, the Department of Citizenship and Immigration Canada (CIC) issued a new policy on the detention of non-citizens. Due in large part to growing concerns regarding the costs associated with detention, the new policy sought to limit the numbers of people in detention, and hence the costs entailed by detention, by focusing the attention and energies of immigration officials on criminality as the primary reason for detention. Detention was to be regarded as a last resort, justified only when there was a real possibility that release would endanger the Canadian public. The policy also sought to facilitate the release of people already in detention by setting dollar figure amounts on the sureties/bonds sufficient to obtain a release. The strongest opposition to this policy emanated from immigration officials who reacted with considerable hostility to the proposals. They felt that the policy fettered the discretion which they argue is essential for them to be able to effectively and efficiently carry out their duties. Due in large part to the ardent opposition from immigration officials which this policy elicited.

\textsuperscript{70}Transcript of the Standing Committee on Citizenship and Immigration. Tuesday March 31, 1998:3
CIC has since been engaged in drafting yet another new detention policy which is yet to be released. A senior CIC policy official indicated that the issuing of the revised guidelines was an extremely sensitive matter, due in large part to the sensitivity of immigration officers surrounding the question of discretion. The draft copy of the policy which I had acquired was still largely unchanged, with the interesting exception that the Department had taken out the final section of the policy which was entitled ‘Expected Results’. The official confirmed that this section was deleted due to the likely apprehension that it represented an unjustified limit on the officer’s discretion.  

Clearly, discretion represents an integral part of both the professional identity and responsibilities of immigration officers (similar in many respects to the importance of discretion to police officers). In official policy, discretion is commonly represented as the making of ‘good’ or ‘professional’ judgements based on the ‘objective’ evaluation of complex issues of law, fact and relevant jurisprudence. The primary justification for the issuing of departmental guidelines regarding discretionary decision making is the need for greater ‘consistency’ of decisions in accordance with liberal notions of due process and the principles of fundamental justice. In the view of critics, this lack of consistency amounts at best to ‘arbitrariness’ and at worst to racism and other forms of discrimination.

Interestingly, different agents involved in detention and removal decision-making also have differing understandings of discretionary power. Policy officials tend to adhere to a strictly legal Dworkinian view of discretion. In this view, discretion does not exist unless there are ‘options’ laid out in the relevant legislation and the officer has a legislated ‘choice’ as to whether or not to exercise her authority. In this legalistic view, neither customs officials (who make the first referral to immigration) nor immigration officers carrying out the secondary inspection and who may recommend detention. have

---


72 CIC Draft Detention Guidelines, 1996
any discretion: in this view, if certain 'facts' exist, both customs and primary inspection immigration officers must exercise their legislated authority. It is only the Senior Immigration Officer (SIO) who reviews the initial decision to detain who is viewed as having a broad scope of discretion because their decision is not legally tied to the two legislated grounds for detention and because several options with respect to release are built into the legislation. The SIO may detain or release and may impose whatever terms and/or conditions on release that she deems warranted. Curiously, in this view, adjudicators who review the detention decision after 7 days and every 30 days after that are not regarded as having much discretion: as put by one senior immigration official. "At a detention review there really isn't discretion....the adjudicator can only release when they are satisfied the grounds [for detention] don't exist. Or there really is a third element. they [the grounds for detention] can be superseded by the legality of jurisprudence."\(^7^3\)

This official legalistic view of discretion as created and bound by law, differs from a more interpretive view held by advocates and even some decision-makers. This latter view regards discretion, rather than the law, expansively. Discretion here is regarded in relation to the interpretive dimensions of applying general legal rules to individual cases. For example, a former adjudicator with CIC explained that adjudicators have very broad discretionary powers; in making their decision they must interpret both law and policy and the individual 'facts', situations, circumstances, characters, etc., which together all contribute to the assessment of risk (of danger and of flight) which ultimately grounds their decision in law.\(^7^4\) Advocates and non-governmental commentators tend to share this more expansive, interpretive view of discretion.

Moreover, while policy officials maintain that frontline immigration officials do not have any discretion in making their initial detention recommendation, it would seem

---

\(^7^3\) Interview with Neil Cochrane. Director Case Presentation and Detention. CIC. April 23. 1998

\(^7^4\) Interview with Kathryn Clout. Senior CIC Policy Official and former CIC Adjudicator. April 24. 1998
that the immigration officers themselves disagree. The nature, extent and force of the opposition of Immigration Officers to the proposed 1996 CIC guidelines on detention speaks to their own perception of their discretionary power. As reported in the *Globe and Mail* at the time, Immigration Officers felt the proposed guidelines were a "...blatant attempt by management to fetter the actions of immigration officers." 75 More recently, representatives of the Canada Employment and Immigration Union (CEIU) made a submission to the House of Commons Standing Committee on Citizenship and Immigration which, among other things, took serious issue with the 'fettering' of the discretion of immigration officers by management. As put in their submission, "Officers’ decisions are fettered by management whose decision to over-rule and release detainees is based on cost factors rather than health, safety and security of the Canadian public. There are periodic blitzes of releasing people from at [sic] the Immigration Holding Centre (the Celebrity Inn near Pearson Airport) because the Inn has too many people there." 76

Thus, there is a stark contrast between the impression provided by the conventional liberal legal view that discretion has clearly demarcated definitional content and parameters, and the presence of differing and often conflicting meanings and perceptions of discretion which vary depending, in this case, on one’s professional location and preoccupations. 77

### VII) Contemporary Policy Discussions and Debates on Detention

In 1997-98, a Parliamentary Standing Committee heard submissions from a wide range of governmental and non-governmental representatives on the issue of immigration


76Canadian Employment and Immigration Union, Public Service Alliance of Canada. Submission to the House of Commons Standing Committee on Citizenship and Immigration. March 25, 1998:2

77While not discussed here, the uses of discretion are also influenced by more mundane organizational dynamics and imperatives.
detention and removal policy and practices as recommended by the Immigration Legislative Review Advisory Group (LRAG) in their report entitled *Not Just Numbers: A Canadian Framework for Future Immigration* \(^7^8\) Both the Report and Recommendations and the representations and submissions in the hearings attest to the degree to which discretion continues to be a critical and politically contentious issue in the administration of the detention and removal related provisions of the Immigration Act.

LRAG's Report and the recommendations made in it are guided by the understanding that greater consistency and accountability of Immigration decision-making is necessary both in the interest of fairness and in order to restore public confidence in the legitimacy of the system. This understanding underpins the reports' recommendations that the laws, policies and criteria which guide decision-making in this area be clear and transparent. LRAG's recommendations are designed to remedy what the authors term a crisis of public confidence in the Department's ability to enforce the Act. As put by the authors:

> ...immigration and protection laws (should) be implemented in such a way that the criteria used to make decisions are derived directly from the Acts and Regulations, and are readily available to the public in a language they can understand. Transparency is a necessary condition for the effectiveness and efficiency of the immigration and protection programs and the maintenance or public confidence in their administration\(^7^9\)

The discretionary nature of detention and release decision-making is of particular concern. The report observes that:

> The Act currently provides two bases for detention: unlikely to appear for immigration proceedings and danger to the public. The power to detain is vested in immigration officers and adjudicators. With such broad powers given to officers to deprive persons of their liberty, concern has been expressed by many that in the absence of specificity in the Act, and with

\(^7^8\) Legislative Review Advisory Group (LRAG). Ottawa: Minister of Public Works and Government Services Canada, 1997

\(^7^9\) *Not Just Numbers*, 1997:13
virtually no policy guidelines issued by the department. the standards for detention are not transparent and vary greatly from office to office and officer to officer.\textsuperscript{80}

The report's recommendations are designed in large part to limit and constrain the discretion of decision-makers through the development of specific, clear and coherent law and policy. Decisions must not be, or appear to be, arbitrary. The perception of arbitrary decision-making undermines public confidence in the system and thus undermines the legitimacy of the system. The solution, as proposed by the authors of this report, is to develop and implement more rules and clear criteria to guide decision-makers.

The report also recommends that the current Immigration and Refugee Board (IRB), an independent adjudicative tribunal, be replaced by a "Protection Agency" composed of civil servants who will take over the decision-making responsibilities of the IRB. The creation of a single protection agency, the authors of the report argue, will facilitate: the streamlining of procedures; the coordination and integration of the system; the sharing of information; and the development of a common information database. The report acknowledges that the creation of a protection agency staffed by civil servants is likely to raise concerns regarding the legal requirement to ensure the independence of decision-makers. The authors acknowledge that the elimination of order in council appointments and of the quasi-judicial setting of the IRB "might be perceived" as threatening the independence of decision-makers. The authors response to this concern is to state simply that:

...[W]e believe that it is possible to design a protection agency that can be independent as well as sensitive to broader national imperatives. Our vision of the protection agency is that of a structure designed to ensure that its decision makers are free from bias and improper influence as they deal with individual requests placed before them.\textsuperscript{81}

This proposal represents a radical shift away from the thinking which underpinned

\textsuperscript{80}Ibid.. 1997:104

\textsuperscript{81}Ibid.. 1997:84
the creation of the IRB in 1989 under Bill C-55. The IRB was created largely in response to the major, precedent-setting Singh decision by the Supreme Court in 1985.\(^82\) This decision asserted that where issues of credibility are being determined, fundamental justice requires that this determination be made on the basis of an oral hearing. At the time of Singh, the determination of refugee claims was, for the most part, a ‘paper’ determination. There were provisions for hearings but these were not mandatory, and hearings were in fact quite rare. The Canadian government responded to Singh with Bill C-55. It created a refugee determination process with mandatory hearings in every case to be carried out by an independent adjudicative tribunal, the IRB. Bill C-55 also extended and applied due process and procedural rights to claimants. This creation of an independent quasi-judicial adjudicative tribunal was regarded as necessary to ensure and protect the procedural rights of claimants. Essentially, Singh sought to curb administrative discretion in deciding refugee cases. Oral hearings represent a check against ‘arbitrary’ decision-making which was found to be contrary to the Charter, for citizens and non-citizens alike.

A further important dimension of the Singh decision was the care taken by the Supreme Court to ensure that the human rights entrenched in the 1982 Canadian Charter applied to refugee claimants. It ruled that the ‘everyone’ in s.7 of the Charter included not only everyone ‘physically present in Canada’ but also anyone ‘seeking admission at a port of entry’.\(^83\) In effect, this decision extended Charter rights to non-citizens on Canadian soil.

Detractors of the current refugee determination system, such as the authors of the LRAG report, are quick to point out that the Singh decision did not mandate the creation of the IRB and the current determination process, but rather, mandated more narrowly the opportunity for people to state their case and know the case against them: as stated in the LRAG report:

\(^82\)Singh et al. V. Canada (Minister of Employment and Immigration) [1985] 1 S.C.R.177

\(^83\)Re Singh V Canada. 1985: 456-463
The Court recognized that although the absence of a hearing was not necessarily inconsistent with the principles of fundamental justice in every case, the concern with any procedural scheme was not over the absence of a hearing in and of itself, but over the adequacy of the opportunity the scheme provided for persons to state their case and know the case they had to meet.\textsuperscript{84}

It is remarkable that while the creation of the IRB and the convention refugee determination process injected due process and procedural rights into the process, these initiatives was almost simultaneously accompanied by substantive measures which sought to restrict access to the process (such as visa requirements and fresh obligations on airlines) that restricted the numbers of non-citizens who could reach Canadian soil.\textsuperscript{85} Mandel suggests that the government's intention was clearly to limit the practical consequences of \textit{Singh}:

The problem with \textit{Singh} is not that it required hearings for refugee applicants, but the fact that this requirement existed, not "in the air," but in the concrete context of a political determination to keep all forms of immigration subordinated to local and international power relations. In this social context, the purely formal right to a hearing could only legitimate a refugee policy that is the farthest thing from humanitarianism. In this respect \textit{Singh} is exactly like the decisions on criminal procedure. It looks fine when it is detached from its context. The problem is that it cannot be detached from context...And the government has understood that if only it dresses things up in the formal requirements of due process and fundamental justice, it can get away with whatever it wants in the way of refugee policy.\textsuperscript{86}

It is also significant that more recently the Supreme Court of Canada has effectively undermined the guiding principles of \textit{Singh}, leading some legal critics to charge that "...the alien's rights at common law were better than they are now under the..."

\textsuperscript{84} \textit{Not Just Numbers}. 1997:79

\textsuperscript{85}For a more detailed discussion of this issue see Michael Mandel. \textit{The Charter of Rights and the Legalization of Politics in Canada} Toronto: Wall and Thompson. 1989:172-183

\textsuperscript{86}Ibid., p.182
Charter. As these authors describe, in the case of *Dehghani v. Canada*, the Supreme Court ruled that "...a person "held" at the port of entry and forced to answer questions under the threat of criminal charge if he refused, was not "detained" for the purposes of right to counsel under s.10(b) of the Charter." 89

What then was the intention of the LRAG`s proposal that the IRB be abolished and a new protective agency be created? What do these proposals say about the current social and political climate relating to immigration and refugee determination? And what do LRAG`s proposals imply with regard to the relation between law and discretion? It is surely reasonable to suggest that LRAG`s proposals reflect a heightened social and political hostility towards immigrants and refugees which now allows for law and policy proposals which no longer need to be `dressed up` by due process in order to be acceptable to the public. While arguably there has never been overwhelming support for new immigrants and refugees, recent studies report a general "hardening" of public attitudes towards immigrants and refugees over the last few decades. As put in a recent comparative review of public opinion and public opinion polls on immigrants and refugees in Canada and Australia, "The overall conclusion from this scan of attitudes toward immigrants and refugees is that a deterioration has taken place in levels of public acceptability." 90 Moreover, the `streamlining` of the processes through the creation of a central protection agency is consistent with current political and economic preoccupations with cost and efficiency.

---


89 Weinreb and Galati. 1996:3

One can't help but wonder just whose 'protection' is being sought. LRAG's proposals focus less on the primary importance of due process and procedural rights and more on the need to deter unwarranted or 'fraudulent' claims. The contemporary discursive conflation of new immigrants and refugees with criminality, documented and discussed in the remaining chapters of this thesis, has arguably undermined the efforts of those who work on behalf of new immigrants and refugees and who seek to protect their human rights.

Not surprisingly, the debates surrounding the issue of discretion are reminiscent of those heard in the context of any criminal justice policy which has sought to limit the discretion of criminal justice officials. Immigration officials are vigorously protective of their discretionary powers, while non-governmental representatives and advocates argue for the need to limit, regulate and monitor discretion, relying very largely on recourse to law and legal rules. In this contemporary debate, the state's obligation to respect the Charter rights of all people on Canadian soil, citizens and non-citizens alike, is discursively pitted against the State's obligation to protect its citizens from danger. The broad powers of detention accorded to immigration officers and adjudicators are largely justified on this latter ground. Indeed, heightened fears about crime and criminality, a growing preoccupation with 'risk/danger' and the contemporary discursive conflation of 'criminal' and 'immigrant', all contribute to an immigration department that increasingly is governed by, and governs through, 'crime'. As put rather explicitly by Lucienne Robillard, Minister of Immigration in 1998:

I have spoken about finding ways to facilitate the movement and the integration of people. That is an important part of what we do. But...it is only part of what we do, and at CIC we also have a clear responsibility to protect the safety, security and well-being of Canadians. There are criminals and other undesirables who would like to come to a prosperous country like Canada.\(^9\)

Robillard continued to note that the CIC endeavours to balance Canada's "humanitarian

---

9 Transcript of the Standing Committee on Citizenship and Immigration. Wednesday April 29. 1998:23
approach with the integrity of the system and the protection of Canadians.\textsuperscript{92}

VIII) Discretion as Arbitrary and Discriminatory

In contrast to official justifications of discretionary decision-making and the uses of coercive state power to detain and deport non-citizens, critics of the system tend to allege that the broad scope for discretionary decision-making under the Act allows for the widespread making of 'arbitrary' decision-making. The dominant discourses accept without question the binary that pits 'arbitrary' decision-making against liberal ideals of the rule of law. This focus leads almost inevitably to the view that the solution is a legalistic one, achieved through the adoption and application of clear rules and an enforced commitment to the rule of law and due process.

The Oxford Dictionary defines arbitrariness as: "derived from mere opinion: not based on law: discretionary: capricious; despotic." The conventional understanding of arbitrariness is one that pits it in dichotomous opposition to law, revealing again the dominance of the liberal legal paradigm. As indicated in the above definition, it is simply taken for granted that law is not arbitrary by definition. Arbitrary decision-making is contrary to the liberal ideal of the rule of law; it undermines the premise that different judges hearing the same case would arrive at the same decision. Most of the critics of discretionary immigration decision-making centre on this problem of arbitrariness. This focus leads almost inevitably to the view that the solution is a legalistic one: the problem of arbitrariness must be addressed through the adoption and application of clear rules and an enforced commitment to the rule of law and due process.

Many critics also emphasise the need for an external review mechanism of detention decision-making, in addition to adequate judicial appeal. As put by the Toronto Refugee Affairs Council (TRAC), "Because of the high level of discernment and decision-making required in this very difficult area of Immigration work, transparency

\textsuperscript{92}Ibid..p.23
and accountability to an outside body is essential." David Matas, a prominent immigration lawyer and advocate, echoed this recommendation in his brief to the Committee:

Generally, in Canada, police enforcement systems have some form of civilian oversight and civilian-run complaints redress or ombudsman. Immigration enforcement is unique amongst enforcement systems in Canada as a system without any civilian supervision of the policing authorities. The courts with jurisdiction over immigration enforcement officials can correct errors of law, but they can do nothing about bad policies or inappropriate exercise of discretion.

The submission by Mary Jo Leddy, Director of Romero House, a Toronto based voluntary community group which assists refugees, puts the same point in the following way: "...the enforcement section of Immigration Canada operates as a state within a state. It is the only police-like force with the powers of arrest and detention, with the life and death power of deportation, which has no oversight or review committee."

Amnesty International's brief to the Parliamentary Standing Committee on Citizenship and Immigration expresses a similar concern. It is in favour of preserving and expanding legal protections and safeguards for people facing detention in order to guard against arbitrariness. They submitted that "[E]ach case for detention should be tested by a prompt, fair, individual hearing before a judicial or similar authority. This hearing should be automatic and there should be a further review from time to time if detention continues." In their submission to the Standing Committee, the Metro Chinese and Southeast Asian Legal Clinic advocated the rejection of the entire protection agency

---

93 Brief to the Government of Canada about Recommendations by the Legislation Review Committee, Toronto Refugee Affairs Council, March 1998

94 David Matas, "Immigration Removals: A Submission to the House of Commons Standing Committee on Citizenship and Immigration" David Matas, March 1998

95 Comments on Detentions and Removals, Brief to the Standing Committee on Citizenship and Removals, Mary Jo Leddy, Romero House, April 1998

96 Brief to the Parliamentary Standing Committee on Citizenship and Immigration, Amnesty International March 1998
model. The Clinic agrees with the submissions made by the Canadian Council of Refugees relating to the threat posed by the proposed protection agency to the indecency of decision-makers in this field, the unreasonable time lines proposed for the making of a claim and the lack of a true appeal process.

In the view of many advocates and community representatives who work on behalf of new immigrants and refugees, this "arbitrariness" not only leads to inconsistent decision-making, but also allows for the continuing influence of racist and otherwise discriminatory views. It is not insignificant, according to many critics, that the vast majority of people being detained by Immigration are non-white. Allegations of racism in the context of the administration of Canadian immigration law and policy are not new. The historical legacy of an explicitly racist and discriminatory immigration system, it is argued, continues to inflect immigration policy and decision-making despite official declarations to the contrary, despite the removal of explicit discriminatory provisions from the Act and despite the legal requirements of equality and non-discrimination which have been included in the Act and which are enshrined in the Charter. Allegations of racism are notoriously difficult to 'prove' and are often presented in rather indirect and suggestive ways. Specific references to 'racism' are fairly infrequent in official settings like that of the Standing Committee hearings. The Metro Toronto Chinese and Southeast Asian Legal Clinic has been more explicit than most on this question. In commenting on the 1997 hunger strike by people in detention at the Metro West Detention Centre, the clinic's brief observes "..The detainees were striking in protest of arbitrary and indefinite detention as well as inhumane detention conditions. Almost all of the detainees are immigrants and refugees of colour." The brief goes in to some detail on the links between the detention of people of colour "...and rising anti-immigrant and anti-refugee sentiment in recent years."97

The discriminatory historical development and application of Canadian immigration law and policy has been widely studied. There is no question that the degree

---

97 "Submission to the Standing Committee on Citizenship and Immigration..." Metro Toronto Chinese and Southeast Asian Legal Clinic, April. 1998:1
of discretionary powers granted to immigration officers up to and including the 1952 legislation far exceeds the degree of discretionary powers under the current Act. The 1970 Gana decision provides an interesting discussion of the evolution of Canadian immigration law and policy. With respect to discretion, the Gana case observes that the 1952 Act (and those which preceded it) provided that the only people entitled to enter Canada as of right were Canadian citizens and persons having a Canadian domicile. All others required the permission of the Minister acting through his departmental officials. Any decision made by immigration officers regarding admission to Canada was an administrative discretionary decision, not subject to review by anyone other than the Minister. The broad discretion of early immigration administrators was often used in explicitly arbitrary and was employed in explicitly discriminatory ways to exclude 'offensive' or 'undesirable' groups of people (blacks, Chinese, communists, immoral women etc.).

Despite the considerable changes to the law and policy governing Canadian immigration since the Act since the early 1900s, the nature of the concerns and criticisms regarding immigration law and policy in Canada do not appear to have changed much. In her study of deportation from Canada between 1900-1935, Barbara Roberts observes that in the early part of this century the administration of Immigration law and policy was characterised by arbitrariness; by wide and unchecked discretionary powers; by the

---


99The historical discriminatory exclusions carried out under Canadian immigration law and policy will be explored in more detail in a later chapter.

100A communication from Minister Calder in 1919 to the Canadian Jewish Congress could easily be mistaken for a present day communication. It read: “The Act...undoubtedly places large discretionary powers in the hands of the executive and its administrative officers...I need scarcely to assure you that every effort will be made to see that these powers are exercised sanely and reasonably.” quoted in Barbara Roberts Whence They Came: Deportation From Canada 1900-1935. Ottawa: University of Ottawa Press. 1988:198
relative absence of judicial review: and by a startling degree of public and political ignorance about the system. Decision-making relating to detention and deportation were then considered to be 'purely administrative proceedings' which had nothing whatsoever to do with punishment. Even though detention and release proceedings are now 'quasi-judicial' rather than 'purely administrative' the adherence to the principle that immigration detention is 'non-punitive' remains in force. Because detention and deportation were not considered to be criminal proceedings, people subject to these proceedings were not entitled to the same legal protections as an alleged criminal. The same objections are frequently made in the present day. For example, the Canadian Council for Refugees submitted in 1990, again in 1994 and again in 1998 that:

The wide area of discretion left to immigration officers will inevitably lead to some rather arbitrary decisions. Furthermore...while the legal standard of proof to justify detention on the statutory criteria requires that the decision-maker be satisfied on the balance of probabilities that detention is warranted, there are innumerable problems in practice. The degree of evidence required to meet this standard and the nature of evidence presented pose continual problems, leading to detentions based on hearsay and unsubstantiated allegations which are not susceptible to proof or to cross examination. While in criminal proceedings where liberty is in jeopardy the tribunal is bound by the legal rules of evidence, immigration decision makers may act on any evidence considered credible or trustworthy by them.101

The persistence of the 'problem' of discretion in the present day, despite the proliferation of laws and rules which were designed to rationalise, legalise and check the abuses of discretion, suggests that legalistic solutions do not actually represent a real or potential remedy for the problems conventionally associated with discretion in this field. Nevertheless, concerns about 'arbitrariness' issuing out of discretionary decision-making powers issue come from across the whole range of the political spectrum. On one side, arbitrary decision-making (seen as an expression of misplaced humanitarian compassion) is faulted for undermining the necessary 'toughness' of the system which must protect the

Canadian public from dangerous non-citizens. On the other side, arbitrary decision-making (seen as discriminatory) is regarded as a vehicle for unjust, inconsistent and capricious decision-making which penalizes one of the most vulnerable and powerless groups in society, non-citizens. Despite these competing perspectives, the recommended solution to the 'problem' of discretion and its relationship to so-called arbitrary decision-making is the formulation and application of further rigid, specific and clear rules to guide and check its operation.

While the debate is now well known in the context of Canadian criminal justice administration, the debate is less well known, though none the less pointed, in the context of the administration of the Canadian Immigration Act. The underlying assumption guiding policy proposals in both these areas is that discretion exists in the absence of legal rules. Therefore the problems associated with arbitrary discretionary decision-making can be ameliorated by developing more legal rules to guide, constrain and check the uses of discretionary decision-making powers. This is the case despite historical and contemporary experiences which seem to suggest otherwise. A persuasive contemporary example of the limitations associated with the liberal legal presumption of the effectiveness of 'more rules' to regulate the inappropriate uses of discretion is evident in the adoption of pro-arrest policies in cases of wife assault. The very same assumptions underpinned the arguments made in favour of these pro-arrest policies. Police discretion was regarded as the vehicle for sexist and outdated views about gender roles and male/female relations. The solution, it was argued, was to curtail their discretion through the adoption of more rules. Unfortunately, the pro-arrest policies have arguably had many negative 'unintended consequences', not the least of which includes the reality that pro-arrest policies may actually deter certain marginalized groups of women from calling the police rather than deterring the abusers from abusing.\(^{102}\)

---

\(^{102}\) For a recent critical collection of articles dealing with this criminal justice policy dilemma see M. Valverde et al (eds) *Wife Assault and the Criminal Justice System*. Toronto: Centre of Criminology, University of Toronto, 1995.
IX) Conclusion

It can thus be seen how the degree of discretionary power which characterises administrative regulatory frameworks is generally set against the application of and commitment to the rule of law under liberalism. From a conventional liberal perspective, discretion in administrative decision-making has been defended as a crucial feature of individualised justice: that is, discretion allows for the tailored application of general rules and laws to individual cases. It thus provides the flexibility necessary for the humane consideration of the specific circumstances of individual cases. The critical, but still conventional view regards discretion as a vehicle for arbitrariness, tyranny, caprice and discrimination. As such, many have argued for the need to eliminate, or at least constrain and structure, discretion through the application of rules.

The debate has thus been polarized: discretion is good, discretion is bad; discretion is arbitrary, discretion is individualised justice; discretion is needs to be preserved and defended, discretion needs to be curtailed and controlled. The limited framework and circular nature of the debate is the result of the false discretion/law binary on which it rests. This binary implies a clear distinction between the two and that they are inversely related: 'where law ends, discretion begins'. The dominance of the liberal legal paradigm is further evident in the primary place accorded to 'law' as the most important regulatory instrument in society, leading, in this context, to the assumption that the problems of discretion can only be effectively addressed through the constraints and checks offered by law.

This discussion has sought to highlight the problems associated with the conventional understanding of discretion and to support and reinforce a new way of thinking about discretion. It suggests that the largely taken for granted law/discretion binary must be undermined and the primacy of the liberal legal paradigm out of which the binary issues must be displaced. Also necessary is critical scrutiny of the assumption of individual autonomous subjectivity which the conventional understanding of discretion presupposes. This is, admittedly, no small challenge. Discretion has long been associated with the idea of the individual freedom of those to whom discretion has been delegated to
choose, unconstrained by the certainty of legal rules. To begin to think of it differently necessitates the shedding of longstanding liberal assumptions about law and administration and about choice and autonomy. A radical shift in analytical approach is required. In the first instance, this requires that the assumptions that discretion is a residual legal category and procedural checks on discretion necessarily result in greater substantive fairness. It also requires that: the political, economic and social context of discretionary decision-making be taken absolutely seriously: that particular attention be paid to the policies, processes, practices and effects of discretion in a given field of administrative decision-making; and that attention be paid to the diverse locations and junctures at which discretion is employed, both at and within governmental and non-governmental levels.

To view discretion as a form of power promises to provide a way out of the dichotomous impasse imposed by the law/discretion binary. This view requires that we ask different questions about discretion. Rather than asking how can it be eliminated, curtailed or expanded, made more fair or just, the guiding question becomes how does this power work? What are the practical purposes of its use? What are the historically specific discursive processes which inflect this use? What are the organizational and institutional networks and relationships that facilitate the uses of discretion in a particular governmental context? What are the dominant discursive influences on, and the effects of, the uses of discretionary power in the governance of both Immigration and of new immigrants and refugees?

It is these questions that must guide contemporary studies of administrative discretion, and indeed it is these questions that, in part, motivate and guide the examination of the historical development of exclusionary Canadian immigration law and policy in the next chapter.
Chapter Three

Discretionary Power and Discursive Contests: Transitions in the Governance of Exclusionary Canadian Immigration Policy (1947-1976)

I) Introduction

Exclusionary Canadian immigration law and policy has always been underpinned by shifting, historically-specific constructions of who or what poses, or is likely to pose, a threat to the nation and its citizens: who or what the collective Canadian “we” need to be protected from. These constructions are constituted by and mobilized through historically specific, dynamic and analytically distinct discursive rationales which express dominant national insecurities vis-à-vis perceived and conceived external “threats”. Different eras have been dominated by different perceptions of what has constituted the dominant threat. While there is much overlap and intermingling of these rationales, there is discernible the broad outlines of a general shift over the last forty years from a primary preoccupation with governing the perceived threats to the nation’s racial and moral “purity”, to a guiding preoccupation with governing threats to national security; first to its ideological “security”, and then to the “safety” of the nation’s “public”. Taken together with the material edifices which they entail, these preoccupations can be understood as governing logics: sets of discursive rationales that underpin, guide and legitimate, in this case the use of coercive sovereign power against undesirable and/or undeserving non-citizens. As will be demonstrated in this chapter, discretionary power is, and has always been, a central mechanism in the operationalization of these governing rationales.

In this thesis, exclusionary immigration law and policy refers specifically to those provisions of the Canadian Immigration Act which define categories of people who are to be excluded from Canada, either through the enforcement of the “inadmissibility” provisions of the Act, which prohibit the entry of certain categories of people into Canada, or the “removal” provisions which sanction the detention and deportation of
individuals who are already in Canada. Exclusionary immigration law and policy is most obviously an example of the continuing prominence and influence of the ideological and historically enduring conception of state sovereignty. While sovereignty has always provided the foundational rationale for (coercive) exclusions, the conventional understanding of sovereignty cannot encompass the full range of legal and non-legal, historically specific factors which contribute to the mode and manner of governing exclusionary immigration law and policy and new immigrants and refugees.

The existence and exercise of discretion, and in particular of Ministerial discretion, introduces a complicating factor into the view of immigration exclusions as a straightforward example of the application and enforcement of national sovereignty. This Chapter provides an historical map of the transitions in the governance of exclusionary immigration law and policy from the post-second world war period up to and including the adoption of the 1976 Immigration Act. It has two main objectives: first, to demonstrate that the discursive construction of the undesirable and thus excludable non-citizen is historically specific, shifting and dynamic and that the exclusion of these 'undesirables' is facilitated by the mobilization and employment of analytically and historically distinct logics; and second, to demonstrate the central place of (exclusionary) discretionary power which facilitates and gives coercive effect to these logics under a liberal regime of governance.

As explicitly discriminatory grounds for exclusion (race, morality, political ideology and various combinations thereof) were legally, politically and socially delegitimized in a process which began in the mid-1950s and took hold through the 1960s, 'security' emerged as the dominant guiding logic of exclusionary Canadian immigration law and policy, supplemented by increasingly influential concerns about criminality. The guiding currency of ideologically-based security discourses was at first complicated and ultimately diminished by a variety of national and international factors and by historically specific discursive contests. For example, the dissipation of the threat of international communist revolution over the 1980s and the end of the Cold War entailed an increased preoccupation with the threat to national security posed by
international terrorism, as opposed to domestic forms of political subversion which had previously been the dominant concern.

It was during the 1960s that the concepts of 'criminality' and 'dangerousness' began to gain prominence in the official definition of undesirability. They were mobilized and employed in association with increasingly modified conceptions of 'security' in the governance of exclusionary immigration law and policy. This development began to displace the guiding official preoccupation with the threats posed first, by communist revolution and later, by international terrorism as criminality and dangerousness emerged as the governing rationale for exclusionary Canadian immigration law and policy and for a wide range of enforcement-oriented legal and political reforms.

This shift was facilitated by a variety of national and international factors which eroded the force and guiding influence of liberal humanitarian discourses and liberal legal discourses which had raised a potent and sustained challenge to 'illiberal' exclusionary practices over the 50s and 60s. Liberal legal and humanitarian discourses had contributed in a significant way to the implementation of the 'non-discriminatory' point system and the creation of the Immigration Appeal Board in 1967. Their discursive influence can be seen to have peaked in the 1970s, culminating in the passage of the 1976 Act and its creation of an independent administrative and legalised system of refugee determination. These legislative and policy developments have been conventionally represented as the triumph of liberal and progressive values and interests over illiberal ones. And in a sense they were, but it was a short-lived triumph. As will be argued in Chapters 4 and 5, the intrusion of other preoccupations, primarily those related to 'fraud' and 'criminality', soon eroded the dominance of inclusionary liberal discourses.

This chapter maps and details the discursive shifts and contests which shaped the development and application of exclusionary Canadian immigration law and policy, from the extremely discretionary and explicitly discriminatory 1952 Act to the distinctly liberal 1976 Act. During this period, moral and racial national purity discourses and the 'illiberal' procedures of Ministerial discretion were increasingly contested by and through the mobilization of legal and humanitarian liberal discourses. However, significantly, the
use of sweeping Ministerial discretionary powers to exclude on the basis of ‘security’ was not, in any serious way, challenged during this time. Indeed it can be seen to intensify.¹

II) Shifts in the Governance of Exclusionary Canadian Immigration Law, Policy and Practice (1947-1976): Liberal Challenges and Legal Triumphs

What follows is a review of the shifts in the governance of exclusionary Canadian Immigration law and policy beginning with Mackenzie King’s famous 1947 Statement on Immigration and ending with the crowning achievement of liberalism in this field of public policy: the 1976 Immigration Act. In this review, particular attention is paid to the discursive rationales and contests which have inflected and justified the uses of exclusionary ministerial discretion over the years. Ministerial discretion has both exclusionary and inclusionary preoccupations. However, the exclusionary work of Ministerial discretion has a much longer and more entrenched discursive history than the inclusionary work of Ministerial discretion. This analysis seeks to demonstrate that during the period under review, exclusionary Canadian immigration law and policy was guided and justified by two dominant analytically distinct governing rationales: (racial and moral) national ‘purity’ and ‘security’.

There was during this period a shift away from the identity based exclusions entailed by national purity discourses towards ‘risk’ based exclusions (assessing and governing (future) ‘threats’ to national security and public safety). The extent to which this shift has entailed a change in the composition of the actual subjects deemed undesirable and/or undeserving of exclusionary immigration governance is, however, less certain. There is reason to believe from the historical evidence that in this important area.

¹ As argued in Chapter Two, while discretionary power is not necessarily exclusively located in what is conventionally understood as ‘the state’, and while discretionary power is not necessarily created by law, nor does it exist only in relation to law, the examination of the development and application of legally sanctioned discretionary powers of the state to ‘exclude’ does nonetheless provide an important window through which dominant historically specific social concerns and governing rationales in relation to the governance of exclusionary immigration law and policy may be examined.
discursive shifts in the construction and mobilization of undesirability does not necessarily entail any real change in the people being governed and excluded.

While it is beyond the scope of this study to consider in any detail the earlier period of Canadian immigration history, the logic of (racial and moral) purity had long interacted with criminality concerns in the governance of exclusionary immigration law and policy\(^2\). And indeed, the influence of the logic of security had already become increasingly dominant with the emergence of war time insecurities leading to the frequent use of discretionary powers to exclude against 'subversives', 'enemy aliens', communists, anarchists and the like. It should also be stressed that this analysis does not imply that criminality issues had never before played a part in exclusionary Canadian immigration law and policy. The point being made here, and to be substantiated in the following chapters, is that while criminality has always been, in varying degrees, a key ground for the exclusion of 'undesirables', the dominance which it now enjoys in the governance of contemporary law and policy is unprecedented.

The post World War two ebb and flow of the Cold War heightened national insecurities and helped to fortify the logic of security and the Canadian security apparatus. However, the argument being made here is that as the logic of (racial and moral) purity was increasingly undermined in the 1950s, 1960s and early 70s, the logic of 'security' emerged, increasingly supplemented by criminality discourses, as the guiding rationale for coercive sovereign exclusions, despite periodic ebbs in the Cold War and despite liberal legal challenges to the uses of ministerial exclusionary discretion.

a) Mackenzie King's 1947 Statement to Parliament on Immigration

In his speech in 1947, Prime Minister Mackenzie King gave expression to the underlying principles of Canadian immigration law and policy of the day, namely: admission to Canada was a privilege and not a right; immigration law and policy should

---

not exceed the nation's economic 'absorptive capacity'; and that immigration law and policy should not interfere with the demographic 'character' of the nation. The relevant and often quoted section of this speech reads as follows:

The policy of the Government is to foster the growth of the population of Canada by the encouragement of immigration. The Government will seek ...to ensure the careful selection and permanent settlement of such numbers of immigrants as can advantageously be absorbed in our national economy...I wish to make it quite clear that Canada is perfectly within her rights in selecting persons whom we regard as desirable future citizens. It is not a 'fundamental human right' of any alien to enter Canada. It is a privilege. It is a matter of domestic policy...The people of Canada do not wish, as a result of mass immigration, to make any fundamental alteration in the character of our population....Any considerable Oriental immigration would...be certain to give rise to social and economic problems.3

Many have read this excerpt as evidence of 'how much has changed' since 1947. Arguably as well, it speaks to 'how much remains the same'. With the exception of the explicit expression of the racist dimensions of Canadian Immigration law and policy of the day, concerns about Canada's economic 'absorptive capacity' and the sovereign declaration that 'entering Canada is a privilege and not a right' continue to be dominant.4 The assertion of Canada's sovereign right to be selective with respect to whom it allows to enter the country has always underpinned the government's control of immigration law and policy and has always provided the guiding rationale for broad Ministerial exclusionary discretion within this context. This was true then as it is today.

The fact that King's 1947 statement made explicit and racist reference to the potential threat to the 'character of our population' posed by 'mass' immigration, speaks


4 While the phrase 'absorptive capacity' continues to be banded about, immigration law and policy is being increasingly reoriented to a 'long-term' plan rather than the 'tap on, tap off' approach entailed by the notion of 'absorptive capacity'.
to the degree to which racist beliefs, articulated and applied in part through the mobilization and application of national ‘purity’ discourses, had been socially, legally and politically dominant prior to the 1960s. Until the 1960s, racism had intermingled with morality and class in the constitution of the ‘(un)desirable’ citizen: and ‘(un)desirability’ was linked discursively with the need to protect national ‘purity’.5

b) The 1952 Immigration Act

The 1952 Act was the first Immigration Act since 1910. The 1952 Act codified existing racial and moral bases for exclusion and sanctioned sweeping Ministerial discretionary powers to exclude, either through inadmissibility or through deportation. The 1952 Act gave the federal Cabinet sweeping discretion which, "...in practice devolved to the minister responsible and his officials"6 over (in)admissibility and deportation.

i) Discretionary Power and Exclusionary Provisions

Section 61 of the Act gave the Cabinet the power to exclude people from Canada, on the following grounds:

(1) nationality, citizenship, ethnic group, occupation, class or geographical area of origin;
(ii) peculiar customs, habits, modes of life or methods of holding property;
(iii) Unsuitability having regard to the climatic, economic, social, industrial, educational, labour, health, or other conditions or requirements existing temporarily or otherwise, in Canada or in the area or country from or through which such persons come to Canada; or (iv) Probable inability to become readily

5The racist dimensions of the historical development Canadian Immigration law and policy are now well documented. However, the logic of national ‘purity’ was not only about ‘racial purity’. For an interesting analysis of the shifting and dynamic relations between race, morality and class in the historical discursive construction of ‘purity’ in English Canada see Mariana Valverde The Age of Light, Soap and Water: Moral Reform in English Canada. 1885-1925. Toronto: McClelland and Stewart Inc., 1991

6Kelley and Trebilcock 1998:324
assimilated or to assume the duties and responsibilities of Canadian citizenship within a reasonable time after admission.  

This section is virtually unchanged from Section 38 c) of the 1910 Act with the one exception that the 1952 provision substituted ‘ethnic group’ for ‘race’. There is little doubt that these provisions were designed to exclude non-whites. Particularly noteworthy is the inclusion of ‘unsuitability’ having regard to ‘climatic’ conditions. A 1952 letter from Walter Harris, then Immigration Minister, makes the following telling claims:

In the light of experience it would be unrealistic to say that immigrants who have spent the greater part of their life in tropical or sub-tropical countries become readily adapted to the Canadian mode of life which, to no small extent, is determined by climatic conditions. It is a matter of record...that natives of such countries are more apt to break down in health than immigrants from countries where the climate is more akin to that of Canada. It is equally true that, generally speaking, persons from tropical or sub-tropical countries find it more difficult to succeed in the highly competitive Canadian economy.  

As observed by Jakubowski. "From the Immigration Act of 1910 up to and including the Act of 1952, Section 38 c) was the principle instrument through which the implicit ‘White Canada’ policy on Canada was implemented."

This view of the racist history of Canadian Immigration law and policy is today widely admitted and well-established. In a recent comparative analysis of Canadian and Australian immigration and refugee law and policy. this history is succinctly and directly summed up as follows:

____________________________


8 Noteworthy (CCF member) reading letter from Harris. in House of Commons Debates 1953:4351-2, quoted in Kelley and Trebilcock 1998:325

9 Jakubowski. Immigration and the Legalization of Racism, 1997:17
The history of immigration in Australia and Canada is indelibly linked to racism, prejudice and xenophobia. Attitudes of racial prejudice shaped immigration policy in both countries, leading to the "White Australia" and "White Canada" policies which determined immigration selection procedures until the 1960s.\textsuperscript{10}

The list of 'Prohibited Classes' was also virtually unchanged from the 1910 Act. It included "...idiots, imbeciles, morons, epileptics, beggars, as well as the insane, the diseased, and the physically defective ...individuals convicted of crimes of moral turpitude, and those advocating the subversion of democracy." \textsuperscript{11} Also among the prohibited were "...persons seeking entry for immoral purposes...persons who were public charges or judged likely to become such, alcoholics, drug addicts and persons who had trafficked in drugs."\textsuperscript{12}

It is analytically significant that the 1952 Act included the new categories of homosexuals, drug addicts and traffickers. The exclusion of homosexuals was officially justified by, and discursively linked to, cold war 'national security' concerns: 'sexual deviants' represented a national 'security risk'. As argued by Philip Girard, this discursive linkage was in large part the result of pressure from the RCMP and the American security establishment "...which had decisively linked Communism and homosexuality in its collective mind as part of a complex of cold war subversion."\textsuperscript{13} Arguably the exclusion of homosexuals under the 1952 Act, and the ease with which this exclusion was accepted,\textsuperscript{14} speaks, at least in part, to the growing primacy of the logic of security evident in the 1952

\textsuperscript{10} Adelman \textit{et al} (eds) 1994:547(Vol.2)

\textsuperscript{11} Kelley and Trebilcock 1998: 325

\textsuperscript{12} Lorne Foster \textit{Turnstile Immigration: Multiculturalism, Social Order and Social Justice in Canada}. Toronto: Thompson Educational Publishing. 1998: 89

\textsuperscript{13} Cited in Whitaker 1987:37

\textsuperscript{14} Whitaker describes the absence of any official opposition to the clause prohibiting homosexuals and how it "...was rushed through parliament with no discussion whatsoever..." Whitaker 1987:37
Act. Although the exclusion of homosexuals under the logic of security may appear as a new, disguised version of moral exclusions, in this analysis it indicates the intermingling or ‘hybridity’ of logics of governance in this field. Moreover, the addition of exclusion on the grounds of ‘drug addiction’ and ‘drug trafficking’ signals the emergence of an increasing preoccupation with ‘true crimes’ central to the logic of criminality.

With respect to discretion, section 39 of the 1952 Act preserved the very wide scope of discretionary power which had characterised the 1910 legislation. Section 39 ensured that the government, acting through the Minister and his officials, continued to have the ultimate and final authority over exclusionary decisions. It decreed that there was to be no recourse in law for those deemed undesirable and excludable:

No court and no judge or officer thereof has the jurisdiction to review, quash, reverse, restrain or otherwise interfere with any proceeding, decision or order of the Minister, Deputy Minister, Director, Immigration Appeal Board, Special Inquiry Officer or immigration officer had, made or given under the authority and in accordance with the provisions of this Act relating to the detention, deportation of any person, upon any ground whatsoever, unless such person is a Canadian citizen or has Canadian domicile.15

While the 1952 Act had provided for Immigration Appeal Boards, they could only decide on a limited category of deportation decisions and even then only on matters of law, and due to the wide degree of discretion accorded to immigration officials, errors of law were quite rare. Moreover any of their decisions could be overturned by the Minister.16 The 1952 Act “gave the Minister powers of discretion such as to give him

15 Quoted in Hawkins, 1987:103

The 1952 Act had defined and entrenched the legal concept of ‘domicile’ for the purposes of immigration policy. Domicile, defined as a 5 year period of residency in Canada, effectively served as a protection against deportation for non-citizens. Interestingly this ‘protection’ was eliminated from the 1976 legislation.

potentially the last word on every individual case." It was indeed a 'discouraging document'. characterised by a 'negative tone' and a 'heavy emphasis on exclusion'.

ii) 'Security' Concerns and the 1952 Act

It is analytically significant that during the late 1950s and throughout the 1960s, the 1952 Act was increasingly seriously scrutinized and criticized for its sanctioning of exceedingly wide discretionary power to exclude and its explicit racism. Yet the use of similarly broad ministerial discretion for the exclusion of 'security' threats (rather than 'purity' threats) attracted next to no critical attention. Moreover, not only were discretionary powers to exclude on the basis of 'security' immune to any serious criticism, the continued preservation and enforcement of this exclusionary discretion, in the context of the growing challenge posed by liberal legal and humanitarian discourses during the late 1950s and 1960s, was a matter of particular concern. Similarly, while the explicitly racist dimensions of the 1952 Act and the racist parameters of the 'desirable citizen' enforced through the application of ministerial discretion has been a matter of considerable recent critical scholarly attention, the ideologically and politically motivated exclusions affected through the security provisions of the Act have elicited far less critical attention.

The notable exception to this is provided by Reg Whitaker in his carefully researched and detailed study of Canadian immigration policy. Whitaker's work raises a potent challenge to the conventional liberal view of the 1960s as a period of liberalization, legalization and progress in the field of immigration. He highlights the fact that although racist and otherwise discriminatory uses of exceedingly wide 'unchecked' Ministerial discretion to exclude were fiercely attacked on 'legal' grounds, similarly broad tracts of exclusionary discretionary power to exclude on political and ideological

---


17 Hawkins 1987:102

18Ibid.,p.105
grounds in the name of 'security' did not elicit parallel legal concerns. During the late 1950s and early 1960s, as increasing public, political and legal attention was being paid to immigration policy and in particular to the racist and discretionary dimensions thereof. "[N]one of these developments appears to have had any impact on the thinking of the security establishment as reflected in the practice of immigration regulation and control."\(^ {19}\) The logic of 'security', already well entrenched during war-time and over the postwar period, can be seen to expand and intensify, free from scrutiny, at the same time as the logic of (racial) 'purity' was increasingly contested. The 1952 Act radically extended the security related exclusionary provisions:

...when the act was passed, broader legislative authority was given to prevent the entry into Canada of persons associated "at any time" with any group about which there were "reasonable grounds for believing" that they advocated or promoted "subversion by force or other means of democratic government, institutions or processes, as they are understood in Canada."

...or were "likely to engage in or advocate" subversion. Moreover, persons "likely to engage in espionage, sabotage or any other subversive activity" were also to be barred.\(^ {20}\)

The critique of the 1952 'security' provisions bears a striking resemblance to contemporary criticisms of the contemporary 'criminality' provisions ushered in with Bill C-44.\(^ {21}\) This should not be surprising given that the 'security' provisions of the 1952 Act largely parallel the 'criminality' provisions contained in the Bill C-44's recent amendment. Both employ discretion in the work of determining the risk of future 'threats' posed: both lack any clear, full right to judicial review; both refuse the right of rejected applicants to be given the reasons for the decision made; and, both remove the right of permanent residents to appeal. As with the 1995 'danger opinion', under the 1952 Act, with regard to landed immigrants "...the state's invocation of "national security".

\(^{19}\) Whitaker 1987:204

\(^{20}\)Ibid., p.35

\(^{21}\)See Chapter 6 for a discussion of this.
consideration could effectively override any appeal process". Moreover, the definitional content of both the 1952 'security' provisions and of the 1995 'criminality' provisions are vague and unspecified. 'Subversion' and 'likelihood to engage in subversion' were as difficult to define then as 'dangerousness' and 'likelihood of posing a danger' are today. And finally, both provisions are particularly serious in that they operate within the ever present looming sovereign spectre of the threat of deportation.

As with the slippage which is evident in the context of 'purity' discourses with respect to their racist, moralistic and class dimensions, 'security' discourses which facilitated exclusions on ideological and political grounds can also not be separated off from other, then familiar, prejudices. This difficulty in reducing exclusionary logics to a single explanatory axis is acknowledged by Whitaker. He notes that while his primary focus is on the "...political and ideological bias, the illiberality, and the duplicity that...characterised the implementation of Cold War security policies", at the same time, "...some more familiar biases, such as racism, appear anew, cloaked under the rubric of "national security". Similarly, while the contemporary governmental preoccupation with 'criminality' cannot be reduced exclusively to racism, there is no question that constructions of criminality are raced.

While Whitaker considers the exclusionary uses of discretion in the context of the Cold War to be 'illiberal' and thus offensive, I argue that the discretionary power to exclude on the basis of deservedness and/or desirability, whether governed by the logic of purity, security or criminality is distinctively liberal. While legal checks on discretionary decision-making are regarded as practically desirable, they do not represent a substantive

---

22 Ibid. pp.35-36

23 As put by Whitaker with respect to the 1952 security provisions, "...specific criteria for applying these very general terms...were not part of the legislation...but were left entirely to administrative discretion." (Ibid.,p.36)

24 Ibid.,p.9

25 This too is discussed in a later chapter.
challenge to the fundamentally exclusionary preoccupations of discretionary power in the construction and reproduction of dominant conceptions of (un)desirability which guides exclusionary immigration law and policy. Liberalism, as a mode of governing, imposes restrictions on the state’s power to govern individuals through coercion and force. Exclusionary discretionary power, in the context of immigration, is the central mechanism which facilitates and justifies the continued sovereign control of non-citizens under a liberal regime of governance. One might even suggest that legal protections actually facilitate this state sanctioned coercion by setting out how exclusions, which in the case of deportation is extremely coercive, can be carried out - legally.

c) Rising Opposition to the 1952 Act

The two problems of 'unfettered' or 'absolute' discretionary power (unless it was linked with security security issues), and of racial discrimination were the primary focus of increasingly widespread and vocal legal, social and political opposition throughout the 1960s. Most prominent amongst the critics of the 1952 Act were lawyers and members of parliament. The former were predominantly concerned about the absence of legal protections afforded to non-citizens in the context of (non-security related) discretionary exclusions and the 'arbitrariness' that this absence entailed, and the latter were particularly troubled by the discriminatory nature of the existing legislation, a concern largely explained by pressure from increasingly ethnically diverse constituencies.

During this period, liberal legal discourses relating to notions of due process, individual rights and equality gained currency and converged with humanitarian and compassionate discourses relating to notions of social responsibility, social citizenship and international human rights that had emerged and had gained social, political and legal influence after the Second World War; "...even if admission to Canada was still considered a privilege and not a right, basic due process protections were coming to be
seen as properly extended to aliens...the values and the interests that were driving immigration policy had taken on...a much more liberal complexion."

Opposition to the 1952 Act also emanated from economic sectors but for quite different reasons. The Economic Council of Canada, sensitive to the rising unemployment and recession in the late 1950s, was particularly concerned by what it viewed was an excessively large and unwieldy sponsorship program which was issuing in a veritable ‘flood’ of unskilled, dependent immigrants. This concern was fuelled and heightened by "...the increasing need for skilled manpower in Canada and the very real difficulties experienced by the unskilled in the Canadian labour market."

In 1964, the Council went on record with the assertion that "the future prosperity of a nation will depend upon an adequate supply of professional, technical, managerial and other highly skilled manpower." Indeed, it was these two related issues, the impact of the sponsorship program and the argued need for more highly skilled workers that underpinned the government’s 1966 White Paper, to be discussed in more detail below.

Finally, the wide scope of Ministerial discretion under the existing legislation had created an unmanageable workload for the Minister. From an organizational and pragmatic perspective, there was simply “too much discretionary power for the Minister.”

Of particular interest in this analysis are the debates, discussions and reforms which addressed the issues of ministerial discretion and exclusion. As with Bill C-44 in 1995, it was the liberal legal paradigm (individual rights, due process, procedural justice), coupled with related egalitarian (non-discrimination) and humanitarian (protection, asylum) discourses which guided and set the parameters of the sustained critique of the

26 Kelley and Trebilcock 1998:345
27 Hawkins 1972:160
29 Hawkins 1972:103
exclusionary discretionary powers. Indeed, there were as well serious and contentious questions relating to more inclusionary policies such as the need for skilled over unskilled labour, the nature and impact of the sponsorship program and the need for long-term immigration planning v. the ‘tap-on, tap-off’ approach. However, the most severe criticism, which began in the mid-fifties and gained momentum through the 1960s, emanated from lawyers and MPs and focussed on individual rights and non-discrimination in the context of exclusions.\footnote{Ibid., p.103}

Two of the first and most effective critics of the 1952 legislation were E. Davie Fulton, a member of the conservative opposition and John Diefenbaker, leader of the Conservative party in 1956, both lawyers. In 1955, Fulton, on behalf of the Conservative party, initiated a critical debate in the House of Commons on immigration law and policy. Fulton submitted various legal reports and resolutions of the Canadian Bar Association and made numerous critical representations focussing primarily on the issue of individual rights, or the lack thereof, in the context of discretionary ministerial exclusions. The substance of the legal critique of Ministerial discretion under the 1952 legislation parallels that which was levelled at Bill C-44 in 1995. Fulton charged the government with ‘administrative lawlessness’ that denied ‘...simple justice to Canadians and non-Canadians alike.’\footnote{Ibid., p.110} Echoing the political and legal defence raised today with respect to broad ministerial discretionary powers to exclude, the government of the day responded repeatedly that ‘...there were no rights attached to immigration. It was a matter of discretion only and could not be made a matter of law.’\footnote{Ibid., p.110} Fulton expressed the dominant liberal legal view which has remained remarkably consistent over the years. The following quote from the House of Commons Debates in 1955 could be mistaken for a current objection to the 1995 ‘danger opinion’ legislation. Fulton argued.

\footnote{Ibid., p.103}
When reasons for decisions do not have to be given...as the government maintains that they do not. and when ministerial discretion. which means in the nature of things. departmental discretion. is the sole arbiter. then error. corruption. favouritism. and injustice are invited and rights and liberties are denied in principle as well as in fact.  

i ) The 1962 Immigration Regulations

The 1962 regulations mark the beginning of the government’s dismantling of the discriminatory dimensions of the Act and the growing recognition of the need to protect the individual rights of non-citizens. The 1962 Regulations began to respond to the challenges raised by liberal legal discourses. They removed the ‘preferred classes’ provision (P.C. 1956-785) which had proclaimed a hierarchy of desirability based on country of nationality. The 1962 Regulations removed almost all traces of racial discrimination with the principle of ‘skill over ethnicity’(s.31(a)). The notable exception was that racial discrimination was preserved in sponsorship provisions which continued to favour Europeans. While all immigrants could sponsor close relatives. s.31(d) stipulated which countries could sponsor more distant relatives. “this clause ruled out Asia and all of Africa except Egypt.”

The regulations also began to respond to the growing concerns regarding the protection of the individual rights of non-citizens facing deportation. They enlarged the jurisdiction of the existing IAB to allow all people facing deportation the right to appeal. not just the previously selected categories. The Immigration Department claimed in its annual report that under the new regulations. “…the responsibility for the hearing of all appeals rests with the IAB which is completely independent of the Immigration

---


34 Kelley and Trebilcock 1998:328

35 Hawkins 1972:126
This overstated the case. Because the Regulations had been effected through executive order, rather than through an amendment to the Act, the Minister could still reverse any decision rendered by the Board. In addition, the IAB could still not hear sponsorship appeals and the Minister continued to be ‘swamped’ with individual cases.

It was not until the Immigration Appeal Board Act was passed in 1967 that a more independent and effective Board was created.

ii ) The Sedgewick Report of 1965

In 1964 lawyer Joseph Sedgewick was asked by the Minister of Justice to investigate 23 allegations of unlawful detention (20 of these were Greek sailors who had jumped ship in Canada). Sedgewick was also asked more generally to report on the ‘...procedure now being followed in relation to the arrest, deportation and prosecution of persons who enter Canada or remain there illegally.’ Sedgewick was subsequently asked by Prime Minister Pearson to expand his terms of reference to include consideration of ministerial discretion.

Accordingly, Part One of the Sedgewick Report addresses the particular issue of detention and Part Two examines the question of ministerial discretion. Sedgewick’s investigations led him to defend the detention and deportation decisions made by the Immigration Branch in the cases at hand. He concluded that there had been no unlawful detention or unlawful denial of rights in these cases. He added that the problem of deserting seamen ‘...has developed into a wholesale...and deliberately planned method of

\footnote{36 quoted in Hawkins. 1972:126}
\footnote{37 Kelley and Trebilcock. 1998:333}
\footnote{38 Hawkins. 1972:126}
\footnote{39 Ibid. p. 127}
\footnote{40 Ibid., quoted on p. 145}
circumventing Canada's immigration laws." Sedgewick was disturbed by what he perceived as unfair criticism levelled at the Immigration Branch in this matter, stating that "many of the attacks which have been made in this affair have been ill-founded or exaggerated."

At the same time as Sedgewick was preparing his report, the Department of Citizenship and Immigration was preparing its own report on the deportation process as part of a series of background papers in preparation for a Departmental White Paper on immigration. The Department repeated the view that criticisms of the Department with respect to arrests, detention and deportation were largely unfair and groundless. Nevertheless, this report reluctantly conceded that although the criticisms had been ‘unfair’,

...it is felt that so long as the Immigration Act provides such broad and uncontrolled power to detain individuals, the legal profession, the public and the press will tend to be very critical of anything which appears to be an excessive and improper use of this power. Although current immigration policy and procedures have established safeguards against abuses, the public will not be satisfied that these safeguards are genuine unless they are reflected in the legislation itself.

Justice must not only be done, it must be seen to be done. Thus in all contexts, including detention, the guiding preoccupation during this period was the legal implications of ‘unchecked’ ministerial discretion to exclude - not with ‘exclusion’ per se.

The Sedgewick Report is most commonly regarded as a prescription for the curtailment and ‘checking’ of Ministerial discretion and ultimately as a prescription for the revision of the Act. Sedgewick recommended, among other things, that the IAB

---


42 Ibid., p. 38

43 Quoted in Hawkins 1972: 147
should be completely independent and authoritative, subject to a right to appeal to the courts. He also recommended that s.39, which preserved the final discretionary authority of the Minister, should be eliminated and that "...express provision should be made from the IAB to the Exchequer Court of Canada on questions of law with a further appeal to the Supreme Court of Canada with leave of that Court." Sedgewick repeated once again the dictum that entering Canada is a privilege and not a right, a principle which he recommended should be clearly stated in the Act. He firmly believed that due process provisions and the respect of the individual rights of non-citizens should not unsettle or disturb the state's sovereign and absolute right to select desirable citizens and exclude others.

The Sedgewick Report is most often remembered for its liberal legal recommendations relating to the curtailment of ministerial discretion. However for the purpose of this review, Sedgewick's unwavering support for the 'state's right to exclude' is of particular interest. The preservation and justification of this sovereign right rests in his view upon the logic of security: that it was 'both necessary and proper that every effort be made to exclude aliens who are undesirable for security reasons.' The inherently imprecise meaning of 'security reasons' is highlighted by Sedgewick himself. He acknowledges that "...while the word security immediately brings to mind the struggle with communism" it also covered other "totalitarian causes" and criminals as well. The mention of 'criminals' in the context of national 'security' is significant, providing an early signal of the expansion and shifts which the definitional categories of both 'security threats' and 'criminality' would undergo in the years to come. Sedgewick also included several enforcement related recommendations, including a system of 'alien' registration and fingerprinting. Thus, while most have focussed on Sedgewick's recommendations relating to the legalization of Ministerial discretion, it can also be read quite differently: "...[T]he Sedgewick Report can be viewed as a strong restatement of the traditional

\[44\]Ibid., p.148

\[45\]Quoted in Whitaker 1987:220
conservative view on national security and the rights of the state as overriding the rights of individuals especially if they are not citizens." Arguably, the lack of attention to this aspect of the report speaks at least in part to the taken-for-grantedness of the guiding logic of security which increasingly stood alone as the logic of (racial) purity was contested and dismantled.

iii) The White Paper of 1966

At the same time that Sedgewick was carrying out his inquiry, the Immigration Department was preparing its own policy paper on immigration which appeared in 1966. Most commentators regard this document as a fairly predictable document which addressed the dominant concerns of the day; "...an exercise in persuasion for a particular policy." As noted, this report centred most obviously on the need to better manage and control the sponsorship program, the need for increasingly skilled labour (as argued by the Economic Council), the need to substitute ‘skill’ for ‘ethnicity’ in immigration selection procedures, and the need to create a more independent Immigration Appeal Board. The new Board would take over the Minister’s discretionary power to exclude and would be limited in its jurisdiction only "...by the right to appeal its decisions on questions of law to the Supreme Court of Canada with leave of that court."

\[\text{\textsuperscript{51}} \text{White Paper 1966:35}\]

\[\text{\textsuperscript{50}} \text{This view was strongly opposed by representatives of Industry and Labour who argued that the selection of ‘skilled’ over ‘unskilled’ immigrants would ‘hamper frontier development’, and who argued that there was still a need for unskilled immigrants. (Hawkins 1972:162)}\]

\[\text{\textsuperscript{49}} \text{Hawkins 1972:160}\]

\[\text{\textsuperscript{48}} \text{White Paper on Immigration, Ottawa: October 1966}\]

\[\text{\textsuperscript{47}} \text{For a detailed discussion of the Sedgewick Report and its recommendations see Hawkins 1972:145-150}\]

\[\text{\textsuperscript{46}} \text{Ibid..p.221}\]
accede finally to the 1951 *International Convention on the Status of Refugees* and create a domestic system to determine whether individual claimants had a well-founded basis for (were ‘deserving’ of) Canadian protection in the form of refugee status.

However, and more significantly for the purposes at hand, this document reflected the ongoing and increasing desire on the part of the government to exclude undesirables from Canada. It provides an early expression of the shift in rationality away from that which justified immigration exclusions on the basis of status and towards exclusions based on governing associated risks/dangers.

The priorities and practices of exclusionary Canadian immigration law and policy had been increasingly complicated by the emergent challenges of liberal legal discourses and by humanitarian discourses. The *White Paper* can be read as a contribution to the redefinition (reconstruction) of the dominant conception of undesirability, which could no longer be legitimately constituted through the mobilization of moral and racial purity discourses and which had to accord to the parameters set by the liberal legal paradigm.

The *White Paper* addressed the exclusionary preoccupations of the day. The ‘vexed’ question of ‘security’, as it was described by Sedgewick52: the elimination of moralistic grounds for exclusion; the need to penalize transportation companies which brought inadmissible (undesirable) people to Canadian shores; and finally the *White Paper* addressed the need to develop new measures to deal with certain prohibited classes. most notably ‘organized’ criminals and subversives.

While the *White Paper* did recommend the elimination of certain moralistic status-based grounds for exclusion. the manner in which this recommendation was justified is revealing and speaks to the flexibility of official exclusionary rationales and to the emergence of ‘risk-thinking’ in government policy. Significantly, the *White Paper* did not justify its recommendation to revise the definitional content of the prohibited classes by reference to any humanitarian or otherwise ‘liberal’ or ‘progressive’ rationales. but rather it argued that while the exclusion of these categories of prohibited classes was

---

52Ibid., p.16
indeed understandable and desirable, given Canada's legitimate desire to keep out 'misfits', the explicit inclusion of moral and social exclusionary categories was unnecessary. *These same people could be excluded for the same underlying reasons but through the application of other grounds.* The relevant section reads:

Person who are undesirable on moral or social grounds ought to be excluded as immigrants but not necessarily as non-immigrants. Although any such flexibility must be balanced by a compensating provision for their prompt removal should they attempt to remain permanently or give other cause. The homosexual, the beggar or vagrant and the chronic alcoholic are at present specifically prohibited. Though not particularly desirable as immigrants or non-immigrants, such people are not true dangers to the national interest by virtue simply of their personal failings. To the extent that they represent an unacceptable risk because of factors associated with their weakness, they will be exclulable on health, criminal or subversive grounds, or, as public charges. They therefore could safely be deleted from the specific list of prohibited classes.  

This is indeed a telling justification. Sedgwick makes no effort to disguise his support for the continued exclusion of these 'misfits' and 'undesirables'. Official construction(s) of desirability as defined in legislation do not necessarily coincide with underlying exclusionary preoccupations, moreover. Official definitions of undesirability may be specifically designed and employed in surreptitious ways which belie the official definitions. Additionally, the categories of 'security' and 'criminality' are revealed to be far less definitive and straightforward than official discourses would suggest.

Ministerial discretion to exclude on 'security grounds' was largely immune from sustained substantive criticism. However liberal legal discourses had begun to draw at least some critical attention to the procedural 'secrecy' of these exclusions, and in particular to the powers of the RCMP with respect to security screening. In 1957, the RCMP's involvement in this national service was raised in the House of Commons in defence of its 'illiberal' security screening and decision-making procedures and processes. In response to the suggestion that non-citizens who were subject to security proceedings should be accorded more individual rights, the response was. "[I]n the nature

53Ibid. p. 25
of things, unfortunately, we cannot prove these things [security reports provided by the RCMP, military intelligence, and the civil service]... There has to be an assumption that the people who do it, like our RCMP... are reliable." 

The question of security and security screening in particular was indeed "a vexed problem", and was clearly not easily amenable to the challenges raised by liberal legal discourses: in the contest between liberal legal and humanitarian discourses and 'security' concerns, national security wins. Thus, while there was an internal departmental review in the late 1950s and a Royal Commission on Security (The 'Mackenzie Commission' of 1969) which coincided generally with a 'thawing' of the Cold War, "Neither of these did more than scratch the surface of the immigration security process; some never even scratched [the surface], despite the growing disillusionment of Canadians with the old Cold War mould and despite some severe external shocks to the old thinking." 55 These 'external shocks' included:

...the collapse of the international Communist monolith of years past, the Sino-Soviet split and its divisive ramifications throughout the left worldwide, the influx of American political refugees during the Vietnam War, and the challenge of left wing refugees from right wing totalitarian violence. 56

Nonetheless, evidence of the mounting challenge posed by liberal legalism is found in the White Paper in its important concession that the current legal definition of subversion was unlawfully broad; "[I]t is important that recognition be given to the fact that the holding or expression of unpopular opinions, or sympathy with such opinion, is not in itself indicative of subversive activity." 57 However, beyond this, the question of the

\[\text{References:}\]

54 Quoted in Whitaker 1987:206

55 Whitaker 1987:204

56 Ibid., p.204

57 White Paper, 1966:24-25
processes and practices relating to security screening was left largely alone in the *White Paper*.  

Following the appearance of the *Sedgewick Report* and the *White Paper*, a Special Joint Committee of the Senate and the House of Commons on Immigration was appointed in 1966 to review, conduct hearings and report on both documents. Even more extensive hearings were carried out in 1975 following the government's release of a Green Paper on Immigration in the previous year as background preparation for the long-awaited new Immigration Act in 1976. In the 1960s and the first half of the 1970s, Canadian immigration law and policy received historically unprecedented critical public and political attention and debate, leading some to describe this period as one which witnessed not only the de-racialization and legalization of immigration processes, but also the democratisation of the immigration policy-making process.  

However, it is surely also significant that as the challenges posed by liberal legal and humanitarian discourses were gaining political and social currency, the number of deportations jumped to a record high. In the 1967-1971 period, deportations jumped to 11,766 (compared to "about 3,500 for each of the preceding five year periods"). Unlike the figures for the depression period, very few deportations fell under the 'public charge' classification. The primary causes were 'stealth or misrepresentation' (57%) and criminality (32%). 870 or 7% were classified as 'other', which "presumably covered deportation for security reasons." While it is assuredly true that the official classifications of causes for deportation reveal very

---

58 The processes and procedures relating to security screening were the focus of increasing political and public attention over the 1960s; however, liberal legal discourses did little to affect existing procedures until the mid-1980s. For more detailed discussion of the specifics of these debates and developments see Whitaker 1987.

59 Kelley and Trebilcock 1998:341

60 Ibid..p.348

61 Ibid..p.348

62 Ibid..p.367
little about the ‘actual’ reasons for deportation and that they ‘...conceal more than they disclose...'63. it is nonetheless analytically interesting. The period currently under review saw the numbers of people forcibly removed from Canadian soil more than triple. Moreover, of the 3,242 people who were deported between 1947/48 and 1953/54, more than half (1,956) were deported for ‘other civil causes’ (security reasons), while the remaining 1,286 people were deported for medical, criminal or economic grounds.64 Thus ‘criminality’ deportations are seen to jump dramatically over the 1960s, while ‘other civil causes’ (‘security’) deportations dropped almost equally dramatically.


During these years, three important pieces of legislation were passed which responded to the challenges raised by liberal legal and humanitarian discourses in the intervening years since Mackenzie King’s 1947 statement on immigration. This section examines these legislative reforms in light of their ‘exclusionary’ dimensions. Together, these legislative reforms conventionally represent the ‘triumph’ of liberalism in the context of Canadian immigration law and policy. However, they also speak to the continuing priority of exclusions and to the guiding influence of the rationale of security which not only continued to justify and underpin discretionary exclusions but which was entrenched and intensified.

Thus liberal legal and humanitarian discourses successfully challenged and ultimately displaced, the guiding logic of moral and racial national purity in national exclusions. They also successfully challenged the ‘illiberal’ discretionary procedures of exclusion. Nevertheless, the nature, substance and social construction of national ‘exclusions’ emerged unscathed; neither discretion, nor exclusion per se was at issue in the liberal challenge: the sovereign right to exclude was taken-for granted as was the idea

63 Ibid., p.367

64 Whitaker 1987:195
that exclusions can be legitimately and ‘objectively’ justified and enforced - with the appropriate legal protections. Thus unlike the explicitly discriminatory and illiberal logic of national purity, which could not possibly be sustained in the face of the liberal legal or humanitarian discursive challenge (and indeed in the face of changing economic and labour market needs and changing international migration patterns), ‘security’ of the nation or public became largely regarded as a reasonably uncomplicated, objective and legitimate justification for the continued preservation, entrenchment and application of wide tracts of exclusionary discretionary power in coercive national exclusions.

i) The Adoption of the ‘Point System’ (1967)

In 1967, the government introduced new immigration regulations which implemented the ‘point system’. The point system is conventionally understood as representing the official adoption of the principle of non-discrimination in the selection of immigrants. The introduction of the point system in 1967 “...officially ended racial and ethnic discrimination in the processing of independent immigrants” by making ‘skills’ the main selection criterion65 thereby ending the implicit ‘White Canada’ policy that had shaped the historical development of Canadian Immigration law and Policy from its inception. However, it should be remembered that the guiding statute was still the racist and exceedingly discriminatory 1952 Act and indeed there is good reason to believe that the introduction of the point system had more to do with pragmatic international and domestic political and economic considerations than principled ones. Canada needed more immigrants and immigration from traditional (white) source countries was declining. 66 Also, as pointed out by Hawkins, there was “...an increasing recognition that Canada could not operate effectively with the United Nations, or in a multi-racial Commonwealth, with the millstone of a racially discriminatory immigration policy

65 Jakubowski 1997:18

66 Kelley and Trebilcock 1998:351
Moreover, critics have persuasively argued that less obvious non-legal discriminatory mechanisms persist: the 'skill' criteria continue to discriminate against immigrants from developing countries; most of the resources committed to the recruitment of migrants continued to reflect a 'preference for British immigrants'; as did the government's placement of foreign visa offices.

ii) The Immigration Appeal Board Act (1967)

As had been recommended in the Sedgewick Report, 1967 also saw the passage of the Immigration Appeal Board Act. It created an independent appeal tribunal which had final authority over all deportation decisions subject to judicial review (with leave) to the superior courts on questions of law and jurisdiction. The creation of the new Immigration Appeal Board (IAB) responded directly to the sustained liberal legal critique of (unchecked) ministerial discretion to exclude which had been gathering momentum since the mid-fifties. The Immigration Appeal Board Act devolved the final discretionary authority over exclusionary decisions from the Minister to the IAB. Unlike the previous and largely ineffective Board, the new IAB could hear appeals on both deportation and sponsorship decisions. Significantly, the jurisdiction of the new Board was expanded. It could now consider not only legal and factual questions but also humanitarian and compassionate factors (known as the Board's 'equitable jurisdiction'). After amendments, the IAB Act provided that the Board could take into account "...the existence of compassionate or humanitarian considerations that in the opinion of the Board warrant the granting of special relief." The liberal legal critique with respect to the issuance of reasons as a requirement of natural justice (the right to know the case made against you) was also satisfied through this legislation which, in its final form, provided that the Board

67Quoted in Jakubowski 1997:18
68Ibid..p.19
69Kelley and Trebilcock 1998:351
70quoted in Hawkins. 1972:163
"...may, and at the request to the parties of the appeal, shall give reasons for its disposition of the appeal."71

Under the IAB Act, the cabinet retained the discretion to specify the classes of relatives that could appeal a negative sponsorship decision - indicating the continued efforts of the government to 'control' the sponsorship program and thus reduce the numbers of predominantly unskilled, dependent applicants. Even more significantly for the purposes at hand was the manner in which the Act dealt with the issue of security exclusions. Opposition members tried unsuccessfully to provide for some measure of judicial review for "...prospective immigrants rejected on security grounds, or on the grounds of an adverse criminal intelligence report, whose cases would not, under the Act, come before the Board if the Minister and the Solicitor General jointly decided that this would be against the national interest."72 The challenges raised by liberal legality with respect to the impact of unchecked discretion on the principles of natural justice lacked the discursive power to unsettle the guiding influence of the logic of security in discretionary immigration exclusions.

Despite these major legal amendments and liberal triumphs in the second half of the 1960s, the 1952 Act was still the guiding immigration statute. While the 1967 Regulations and the Immigration Appeal Board Act were important reforms, ...the flawed 1952 Immigration Act remained the guiding statute and it still contained many of the racist and ethnocentric biases of the early twentieth century, including prohibitions against the entry of a wide range of mental and moral defectives. Clearly, if Canada was serious about immigration reform a new Immigration Act was necessary.73

As further indication of Canada's newly entrenched 'liberal' approach to immigration and settlement issues the government passed the Multiculturalism Act in 1971 which marked a significant shift in official policy long characterised by explicitly

71 Ibid., p.163
72 Ibid., p.164
73 Avery 1995:186
racist and assimilationist objectives. Thus as the 1970s began, liberalism had all but triumphed in the context of Canadian immigration law and policy. All that remained was to develop and implement a new Immigration Act.

iii) The 1976 Immigration Act and the 1978 Regulations

In 1973, the government announced its intention to carry out a major review of immigration law and policy and to publish its findings in a Green Paper prepared by departmental officials and academic consultants. In February of 1975, the Green Paper was tabled in the House of Commons. 21 public hearings were subsequently held across the country by the Joint Committee of the Senate and House of Commons. and in July of the same year the Committee submitted its report and recommendations, most of which were incorporated into the Immigration Act of 1976 (passed in 1977). 74

The 1976 Act is conventionally viewed as representing the culmination of a substantive historical shift in the nature and orientation of Canadian Immigration law and policy: from 'illiberality' to 'liberality'. Consider the following glowing summations made by contemporary political scientists:

The Immigration Act, 1976, as it is known, constituted the most liberal piece of immigration legislation ever to become law in Canada. The Act showed a positive emphasis and set as immigration priorities, the reunification of families, humanitarian and compassionate treatment of

74 While a full discussion and examination of the political and public debates which followed the publication of the Green Paper is beyond the scope of this Chapter, it is worth noting that considerable attention was given by the RCMP and of the Ottawa security establishment to the issue of security exclusions. Others raised the now familiar challenges of liberal legality in the specific context of security: challenges which, as we have seen, do not question the underlying logic of security, but rather which focus on procedural justice and the need to further protect the individual rights of those subject to security related exclusionary processes. These developments are discussed in detail in Whitaker Double Standard. 1987
refugees, and the promotion of programs satisfying Canada's economic, social, demographic and cultural goals.\textsuperscript{75}

Or.

The coming into force of the Immigration Act, 1976 on 10 April 1978 ushered in a new era in the history of Canadian immigration law...the new act was a significant departure from its predecessors.\textsuperscript{76}

Or finally,

...the 1976 Act marked the beginning of a new, more liberal and more cooperative era in Canadian immigration.\textsuperscript{77}

There is no question that the 1976 Act responded favourably to the then dominant discursive challenges raised by liberal legality and humanitarianism. It also and no less centrally entrenched and intensified the guiding exclusionary logic of security to an unprecedented degree.\textsuperscript{78} As observed by Whitaker, one of the few immigration scholars who highlights the continued 'illiberality' of the 1976 Act:

....the drafters of the new Act outdid themselves in offering the government the most sweeping statutory authority to exercise administrative discretion in national security matters.\textsuperscript{79}


\textsuperscript{76}Kelley and Trebilcock. 1998:390

\textsuperscript{77}Hawkins 1988:xv. as quoted in Jakubowski 1997:19

\textsuperscript{78}While this study is primarily concerned with 'exclusions', it is certainly analytically significant that the vast majority of scholarly work which examines the development of Canadian Immigration law and policy focuses on the inclusionary, liberal legal and humanitarian dimensions of the 1976 Immigration Act.

\textsuperscript{79} Whitaker 1987:269
While I share Whitaker's critical concern with discretionary exclusions. I do not regard them as 'illiberal' (except in a strictly legal sense). Rather, the continued and increased centrality of the logic of security (and more recently of criminality) in relation to immigration exclusions and the continued preservation of broad tracts of discretionary power to exclude are regarded as practical and effective technologies of liberal governance. Liberalism as a practical mode of government requires that the use of coercive state powers against private individuals be carefully socially, politically and legally legitimated and checked. From a legal perspective, 'punishment' discourses provide the guiding rationale in the context of the coercive powers of the criminal justice system and 'sovereignty' discourses underpin coercive exclusionary immigration practices. From a social and political perspective, the mobilization of national 'security' discourses and criminality and 'danger' discourses further contribute to the continued effectiveness and legitimacy of the sanctioning of extreme coercive powers in the enforcement of immigration legislation under a liberal regime of government. Under this regime, coercive exclusions in the context of immigration are effected through the operation of discretionary powers. The 'inclusionary', 'humanitarian' and 'egalitarian' dimensions of the 1976 Act are indeed liberal changes and certainly do speak to the relatively newly achieved dominance of liberal discourses as reviewed above. However, the preservation of widely discretionary, exclusionary powers in the Act is no less liberal, in the sense of liberalism as an active form of government, rather than as a relatively straightforward political doctrine or ideology.\(^8\)

The new inadmissibility provisions of the 1976 Act\(^8\), removed all explicit traces of the moral (and racial) grounds for exclusion; gone were the long existing prohibitions against 'physically defective persons', 'homosexuals', 'the insane', 'idiots, imbeciles and


\(^{81}\) This discussion speaks to the provisions as they were set out in the original 1976 Act. There have been numerous and analytically significant amendments to these provisions but these will be taken up in the next chapter.
morons' and the like, and gone was the prohibition against individuals convicted of 'crimes of moral turpitude'. The 1976 Act replaced the 1952 prohibition against persons who were convicted of crimes of moral turpitude with a prohibition against persons convicted of crimes prohibited by Canadian Criminal Law. In their place, s.19 organized and itemized no less than 17 specified grounds of inadmissibility. The breakdown of these 17 specified grounds speaks volumes: one deals with medical inadmissibility of persons due either to the danger posed to public health or safety or the likelihood (on 'reasonable grounds') of causing excessive demands on health or social services (s.19 (1) (a)); another deals with economic inadmissibility due to the existence of 'reasonable grounds' for believing persons to be 'unable or unwilling to support themselves' (s.19 (1) (b); yet another deals with people who are not, in the opinion of an adjudicator, "genuine visitors or immigrants" (s.19 (1) (h); and one deals with those entering Canada without the express consent of the Minister when they were required under the Act (s.55) to have this consent (s.19 (1) (l)). The remaining 13 grounds are entirely devoted to matters relating to criminality and security.

Significantly, the inadmissibility provisions are concerned to exclude not only the 'threats' posed to the nation by subversion, terrorism and/or espionage, threats conventionally linked with the guiding logic of 'security', but also exclude 'threats' posed by 'organized' criminals. S.19 (1) (g) prohibits the entry into Canada of those persons.

...who there are reasonable grounds to believe will engage in acts of violence that would or might endanger the lives or safety of persons in Canada or are members of or are likely to participate in the unlawful activities of an organization that is likely to engage in such acts of violence.83

---

82 The 1976 Act replaced the 1952 prohibition against persons who were convicted of crimes of moral turpitude with a prohibition against persons convicted of crimes prohibited by Canadian Criminal Law. (Kelley and Trebilcock 1998:395)

83 Whitaker 1987:269
As is evident from the above, the inadmissibility provisions of the 1976 Act are both backward and forward looking; concerned with deeds done, deeds ‘likely’ to have been done in the past, and deeds ‘likely’ to be done in the future. The ‘forward looking’ dimension of the provisions "...gives extraordinary leeway to the state since it deals with perceptions of intentions in the future rather than with actual acts in the present or past."\(^84\)

Similarly, with respect to national security, s.19(1) (d) refers to:

persons who have engaged in or who there are reasonable grounds to believe are likely to engage in acts of espionage or subversion against democratic government, institutions or processes, as they are understood in Canada, except persons who, having engaged in such acts have satisfied the Minister that their admission would not be detrimental to the national interest."\(^85\)

S.19 (1) (f) of the 1976 Act deals with “persons who there are reasonable grounds to believe will, while in Canada, engage in or instigate the subversion by force of any government.”\(^86\) The breadth of the discretionary powers being employed, is further increased by the fact that neither “subversion” nor “reasonable grounds” nor “democratic government, institutions or processes” are defined in the Act.

An additional extension of discretionary powers in relation to exclusions is provided by the removal of the concept of domicile in the 1976 Act. In the 1952 Act non-citizens who had been living in Canada for five years were deemed to have acquired domicile and could not be deported, with some exceptions.\(^87\) This deletion "...effectively

\(^{84}\)Ibid..quoted on p.269

\(^{85}\)Ibid..quoted on p.269

\(^{86}\)Ibid..p.269

\(^{87}\)Kelley and Trebilcock 1998:430
extended the threat of removal from Canada to all non-citizens, no matter how long they
had resided in Canada."

Of particular interest is the manner in which the 1976 Act dealt with permanent
residents who were deemed to represent a 'security risk'. This is specifically relevant
given the obvious parallels being drawn between the use of Ministerial discretion to
exclude based on 'security' and that used in the 1990s to exclude on the basis of
'dangerousness' and 'criminality'. The 1910 Act had provided for no appeals against a
deportation order vesting final discretionary authority over admissions and removals in
the Minister. The 1952 Act had created a very limited and largely ineffective Appeal
Board which could only review certain deportation decisions and whose decisions could
be reversed by the Minister. The 1967 Appeal Board Act had for the first time conferred
upon the IAB the authority to stay or quash a deportation order made against a permanent
resident on the basis of "all the circumstances of the case". However, this new power was
subject to the discretion of the Minister and the Solicitor General if they agreed to issue a
certificate indicating that based on security and criminal intelligence reports it would be
against the national interest to provide discretionary relief. Thus, in accordance with the
guiding principle of state sovereignty at common law, discretion was, and indeed still is,
preserved for the Crown to decide whether the continued presence of a non-citizen in
Canada was (not) in accordance with the 'public good'.

The 1976 Act left unchanged the powers of the IAB to hear appeals. Appeal
decisions would consider questions of law, fact, mixed law and fact and compassionate or
'equitable' grounds. However, the Act did limit access to appeal on humanitarian and
compassionate grounds; if the Minister and the Solicitor General issued a jointly signed
certificate that it would be contrary to the national interest to grant such an appeal ('on all
the circumstances of the case'). This certificate is considered 'conclusive proof': no
evidence whatsoever need be filed other than the simple assertion, stated in a

88 Ibid., p.430
certificate, that the minister has “reasonable grounds to believe” that the person is “likely to engage in” certain activities.\(^8^9\)

The original act provided for a hearing before a Special Advisory Board in addition to the filing of a certificate. This was a very odd little board indeed. The Royal Commission on Security had recommended in its 1969 report that a security review board should be introduced as a ‘check’ on this particular use of exclusionary discretionary power. However the Special Advisory Board which was created under the Act was a caricature of the liberal legal discourse that had underpinned its creation.\(^9^0\)

After receiving security and/or criminal intelligence reports from the RCMP, the Minister and the Solicitor General could, if they so chose, make a report to the Board and the Board would give the person in question the ‘opportunity to be heard’. However, even more oddly, the Act also indicated that the person in question could not know the evidence against him and could not cross examine potentially hostile sources of information. This Board only heard one case before it was eventually replaced by the Security Intelligence Review Committee (SIRC) under the *Canadian Security Intelligence Service Act* of 1984.\(^9^1\)

A conventional reading of the significance of the 1976 Act is that it marks the beginning of a new, liberal era in Canadian Immigration law and policy. Such a reading focuses on its inclusionary dimensions and considers, if at all, the exclusionary dimensions as a relatively straightforward and legitimate liberal attempt to ‘balance’ the

---

\(^{8^9}\)Whitaker 1987:271

\(^{9^0}\)Past recommendations by the 1969 Royal Commission to create a Security Review Board to take over responsibility for security and criminal intelligence screening from the RCMP had, not surprisingly, been met with fierce opposition from the RCMP. For a full discussion of this issue and the debates surrounding it see Whitaker 1987.

\(^{9^1}\)Whitaker 1987:271
civil rights of non-citizens with (legitimate) national security and criminality related exclusionary objectives. The forgoing review and analysis of the historical development of exclusionary immigration law and policy suggests a somewhat different interpretation. It is suggested here that the 1976 Act represents the peak of the legislative, political and social discursive dominance of liberal legal and humanitarian discourses in the historical development of Canadian immigration law and policy. Liberal legal principles of due process, individual rights, natural justice and formal legal equality (non-discrimination) were legislatively and procedurally entrenched to an unprecedented degree as was belatedly the humanitarian obligation under international human rights law to protect ('genuine') refugees. Additionally, and for the present analysis more significantly, the 1976 Act represents the legalization, formal entrenchment and expansion of the guiding exclusionary logic of security, supplemented by emerging concerns with criminality, facilitated and enforced by the operation of discretionary power, and justified by reference to traditional notions of national sovereignty.

III) Conclusion

_The construction of a community founded upon the expulsion of textual outlaws...requires the establishment of borders and boundaries. Beyond the boundary can be identified the outlaw, within the boundary exist the members of the community, attempting to 'lead their lives free from fear and in relative security'...Some borders are easily identified: immigration law establishes the nation as a discrete community which is vulnerable to flooding by individuals from other countries._

92 It would be misleading to give the impression that Canada had taken no humanitarian action with respect to refugees prior to 1976. It certainly had. However its actions were largely reactive and _ad hoc._ Canada had resisted the creation of a permanent refugee determination system for 'on-shore' claimants until 1976, having previously preferred to maintain some control over which 'humanitarian' refugee needs should get priority.

What is interesting, as well as somewhat surprising, is the degree to which many scholars of Canadian immigration law and policy either ignore the entire question of ministerial discretion for security and criminality exclusions when discussing the development of Canadian immigration law and policy during the period under review, or treat these issues as a relatively straightforward and necessary way to ensure the 'public good'. They sometimes advocate more individual rights so as to make the process more fair, more just, more legitimate, however the exclusions for certain historically specific reasons are accepted as necessary and right. They do not probe the wider significance of this always and already sovereign and absolute right of the state to use its coercive exclusionary powers against individuals who are not 'legal' citizens. Under a liberal regime of government, such coercive powers must be legitimised, and in the context of immigration, legal notions of 'sovereignty' and political and social notions of the 'public good' have provided longstanding legitimating rationales. While racially, morally or ideologically motivated exclusions are clearly antithetical to liberal legal principles of formal equality and therefore obviously offensive under a liberal regime of governance, exclusions which aim to protect the nation and the nation's public, those governed by the logics of criminality or security, are far easier to justify and are conventionally received as relatively unproblematic and legitimate. Unlike racial, moral or ideological exclusions, criminality and security exclusions are not explicitly discriminatory. They are continuously constituted as neutral, objective and, therefore, legitimate bases for necessary exclusions.

Borders define national boundaries and immigration law and policy governs who will be excluded and how. The powers which are applied are coercive sovereign powers: searching and seizing, arresting, detaining, restraining and forcibly expelling. Under a liberal regime of governance, such extreme uses of coercive state power must be legally justified and socially and politically legitimated; the notions of 'sovereignty' and states as national 'communities' constitute foundational legitimating rationale. Illiberal and
coercive exclusionary state activities are 'liberalised' by legal checks and protections. 
individual rights, appeals, reasons, hearings and such: this makes them less offensive to 
the humanitarian liberal conscience and more congruent with abstract notions of natural 
justice and formal legal equality. However, despite all the legal gymnastics which 
liberalism incites with respect to justifying and legalizing exclusions, the basic processes 
of exclusion continue regardless. And these processes are not legal. They are social and 
political processes, inflected by a variety of discourses and guided by distinctive 
governing preoccupations. Dominant constructions of drug dealers, war criminals, 
terrorists and organized criminals are historically specific social constructions: they are 
raced, gendered, moralised, ideological and classed, less obviously but no less 
significantly than other 'threatening' figures in the past. The "body of crime", or perhaps 
here more appropriately, 'the threatening outsider'. "...is being continually reconfigured 
as feminine, black, young, homosexual, maternal and on and on. Such a process does not 
and cannot end."*4

In its exclusionary work, immigration law and policy produces and coercively 
enforces dominant conceptions of the 'undesirable' Canadian citizen: the 'threatening 
outsider'. This work is not merely repressive and negative. As the undesirable citizen is 
egaged through expulsion, the 'desirable' citizen is produced; as the border is policed 
and enforced, the national community is produced; "Founded upon the illusion of a social 
contract, its inclusionary 'we'...can only exist through the expulsion of the others."*5

Liberalism both complicates and legitimates national exclusions enforced through 
the policing of national boundaries. It requires that exclusions be legal and fair: that 
discretion be checked and constrained. It requires that the state bend over backwards to 
justify and legitimate any use of coercive force against its citizens. While non-citizens are

*4 Ibid.. p.19
*5 Ibid.. pp.12-13
easier to exclude, the state still cannot be seen to be acting arbitrarily: in this sense liberal legality legitimates exclusions. In the context of immigration exclusions the liberal challenge sparked more than a decade of legal and political attempts to justify and legalize the use of discretion in the coercive exclusion of non-citizens. The ideological concept of state sovereignty and its constituent consideration of the 'national interest' or 'legitimate social objectives' provides the foundational legal rationale for 'illiberal' state processes and practices of national exclusion. However, as argued here, national exclusions have to more to do with the social and political processes of constructing and reproducing collectivities than with legal concepts of sovereignty or the 'public good'. 
Chapter Four

Floods, Frauds, and Refugees: The Emergence of the ‘Bogus Refugee’ in the 1980s

I) Introduction

The Immigration Act of 1976 (the Act) has been held to represent a triumph of liberal legal and humanitarian values.\(^1\) Certainly much in the Act justifies this judgment. It curtailed ministerial discretion in sponsorship and deportation appeals. It very substantially removed all explicitly racist provisions in the immigration selection process. It ‘de-raced’ and ‘de-moralized’ exclusionary categories. Most significant for the present analysis, the Act entrenched a permanent, on-shore refugee determination system which embodied in Canadian law Canada’s obligations under the 1951 International Convention Relating to the Status of Refugees and the 1967 Protocol Relating to the Status of Refugees.

Canada has always preferred to select its own refugees. It wants to minimize the numbers of refugees-claimants who succeed in presenting themselves on Canadian soil. The government is much more at ease in the business of selective refugee resettlement, sending teams of immigration officers to refugee camps around the world and hand-picking those they judge most desirable, that is those in good health, able to speak English or French, with the most marketable skills and high ‘establishment potential’. Indeed it wasn’t until 1973 that Canada established an administrative procedure for processing on-shore refugee claims.

Canada had previously admitted large numbers of refugees on a reactive basis through special programs set up to respond to particular international crises. For example, in 1968, one year before Canada finally acceded to the 1951 Convention (a delay attributed to Canada’s fears that it would constrain Canada’s ability to deport on security grounds). Canada responded to the Prague uprising and its tragic consequences by relaxing admission standards and providing transportation grants for Czechoslovakian

\(^1\)See for example Kelley and Trebilcock, 1998; Dirks, 1995, and Dirks 1977
refugees. By 1969, 12,000 refugees had arrived. In 1972 when Ugandan dictator Idi Amin Dada expelled 50,000 East Indians. Canada responded with a special program which sent an immigration team to Kampala to process applications. By 1973 more than 7,000 of these refugees had arrived. The political dimension of this approach to refugee protection became clear in 1973 when Chilean General Augusto Pinochet seized power from the democratically elected Marxist government of Salvador Allende. Despite a brutal crackdown by Pinochet which produced thousands of refugees. Canada did not relax admission policies under a special program: indeed it stepped up security screening for Chilean applicants. Only 780 visas were issued to Chileans by Canada over the six month period following the coup.2

In marked contrast to Canada's comparative readiness to receive refugees whom it has chosen either individually or as specially designated categories of immigrants, responding to the needs of 'spontaneous' self-selected refugees who arrive at Canada's borders had not been part of Canada's history. That the Act introduced an admirable adjudication system for such claimants marks a high point of the impact of liberal ideology on Canadian refugee practices. However, the inclusionary and humanitarian dimensions of this major Act were short-lived. No sooner had the new system been implemented than it was subjected to a variety of potent national and international

challenges which, over the course of the 1980s, effectively undermined the 'inclusionary' and 'humanitarian' dimensions of the legislation and re-asserted exclusionary ones.

It will be argued in this chapter that the guiding inclusionary objective - to identify and protect genuine, and therefore deserving, refugees 'at risk' - was soon re-framed in exclusionary terms. The guiding legislative and policy preoccupation with assessing the risk of persecution that individual refugees would face were they to be returned to their country of origin (and therefore the genuineness of their 'need' for protection) has been increasingly supplanted by a preoccupation to determine, assess and minimize the risks to the Canadian state and society posed by putative 'bogus' refugee claimants.³

During the 1980s, the threatening figure of the undeserving, fraudulent, 'bogus' refugee claimant was born, with the state as the 'victim' of their abuse of the system. The 'fraudulent claimant' became increasingly criminalized as the perpetrator of offences against the state. While the deservedness of genuine refugees continued to be generally accepted, refugee claimants, particularly of 'on-shore' self-selected claimants were increasingly regarded as 'undeserving' and 'undesirable' due to the risk they were seen to pose to the state and to the public through fraudulence and criminality. The convergence and increasing currency of 'fraud' and 'criminality' discourses in the context of refugee determinations, and as will be discussed later in the context also of social welfare, were both guided and strengthened by a more general growing preoccupation with 'victims' evident in most areas of public law and policy. Significantly, while this preoccupation with victims could theoretically have lead to an expansion of sympathy for refugees, it translated into citizens at large regarding themselves as potential victims of criminal immigrants.

The liberal humanitarian 'triumph' of the 1976 Act, and in particular the creation of a national refugee determination system, was not undermined exclusively 'from the outside'. Liberalism itself rests upon very particular notions of deservedness. In the

³ The categories of 'fraudulence' and 'criminality' have also become increasingly linked in social and political discourses with respect to welfare recipients. This development will be examined in considerable detail in Chapter Five.
context of social welfare for example, the deserving recipient of the state's 'compassion' and 'generosity' is an acquisitive, independent, law-abiding citizen who has become needy and dependent 'through no fault of their own': they are a genuine 'victim' of circumstances beyond their control and as such they *deserve* the state's help 'to get back on their feet' again. Their claim to state assistance is well-founded on the basis of their genuine 'victim' status.

These same parameters of deservedness apply in the context of refugee determination. It is only the genuine refugee who, by virtue of their genuine, legally defined, victim status *deserves* Canada's protection and access to the benefits of Canadian citizenship. In both contexts, the 'fraudulent' claimant is the discursive epitome of undeservedness. Add to each of them a suspicion of criminality and an archetypal neo-liberal threat is complete: the 'criminal immigrant' who victimizes the state through fraudulent claims of deservedness in the context of welfare or refugee claims and who poses a risk/danger to the public. 'Inclusionary' welfare liberalism has not therefore been hijacked from the outside. Rather its foundational guiding binaries contain the conditions for its own demise. 'Inclusionary' welfare liberalism is necessarily exclusionary - not everyone deserves to be included, not everyone is equally desired for inclusion. The sorting out of who is (un)deserving and who is (un)desirable is a fundamentally liberal project, just as it is a fundamentally moral one. In this light, the discursive shift in Canadian immigration law and policy over the 1980s and early 1990s. is not accurately represented as a shift from 'liberal' to 'non-liberal' preoccupations. but rather as a shift from welfare liberal to neo-liberal rationales. Thus, it is not simply a question of a shift from liberal to conservative political platforms and agendas. 'Desert' discourses are both inclusionary and exclusionary. While welfare liberalism highlights inclusion, compassion and 'deservedness' and neo-liberalism highlights exclusion, risk and 'undeservedness'

Each represents one pole of the same guiding discourse.

'External' factors do of course come into play. The 'on-shore' refugee determination system was put to the test swiftly after it had been created by the 1976 *Act*. Refugees started arriving 'on shore' in greater numbers and from a wider range of 'non-
traditional' (non-white, poor and non-communist) source countries than ever before. The significant increase in numbers threatened to overwhelm the new refugee determination system, a threat which has continued to govern its operation through the 1980s and up to the present day. Equally important, the Canadian government, though ready to admit more refugees than most other developed countries, remained determined to control their selection and therefore sought to ensure that the number of landed refugee claimants was as low as possible.

Many of the legislative and regulatory developments over this period which were designed to respond to these threats have a distinctly neo-liberal flavour. To manage the threat to the refugee determination process posed by the greater numbers of landed claimants the government sought to streamline and improve case processing time and efficiency (without infringing upon the individual rights of those being processed). In response to a series of severe, poorly managed backlogs, a central preoccupation of all these reforms was to find faster and more efficient ways to weed out blatantly unfounded, fraudulent or otherwise excludable claims at an earlier stage in the process.

However, the threat as it was perceived was not merely numerical. The numbers did indeed impose a significant challenge to the existing system and its staff. However mere numbers in and of themselves were not the subject of the public and political moral outrage and panic to which they gave rise. Rather, people were, people arriving in boats on our shores from 'non-traditional', non-white places with unfamiliar faces making claims to 'our' country's largesse.

During the period under review, legal and policy reforms were justified by reference to the need to contain this threat in order to preserve the 'integrity' of the system and the safety of the public. Less well known but perhaps even more significant was the guiding policy objective to retain control of Canada's borders and of the selection of immigrants. On-shore refugees complicate this governmental objective. No sooner had the Canadian government implemented an on-shore refugee determination system did it proceed to limit access both to the shore and to the system.

This Chapter traces these developments. It begins with a discussion of the nature
and context of the major relevant legislative developments of the 1980s and early 1990s. Particular attention is paid to the changing discursive parameters of exclusion in the context of refugee determination and of immigration law and policy more generally. The increasing influence of the logic of criminality and the mobilization and criminalization of fraud in the governance of exclusionary immigration law policy and practice will be explored in relation to the emergence of the regulatory figure of the ‘bogus refugee’.

II) Reframing Risk in Canadian Immigration and Refugee Law and Policy in the 1980s

Two different conceptions of risk underpin and guide the development and application of Canadian immigration and refugee law and policy. The notion of risk, and the governmental objective to identify, assess and control risk in the context of immigration and refugee law and policy, is largely understood and operationalised in terms of threats or dangers posed by non-citizens seeking access to Canada. They threaten the liberal democratic state, the safety of members of the public, the integrity and efficiency of the state’s administrative systems and the national economy. However, Canada’s obligation under international law to provide protection to refugees means that risk must be assessed in two directions. Canadian authorities must assess as well the risks and dangers which would be faced by individual refugee claimants were they to be returned to their country of origin.

These two contrasting and often conflicting preoccupations with risk/danger have characterized and complicated the development of Canadian immigration and refugee law and policy over the last two decades. It is argued here that during this period, the discursive power of the humanitarian concern to protect those ‘at risk’, has been increasingly supplemented with and ultimately superseded by the governmental preoccupation with excluding those who are thought likely to pose a risk. Changing global conditions, domestic economic insecurity and rising xenophobia and racism account in part for this discursive shift.
i) The Changing International and National Contexts

After the implementation of the 1976 Immigration Act, unprecedented numbers of asylum seekers sought to avail themselves of Canada's protection. Between 1979 and 1990, some 154,000 refugee claims were made in Canada, peaking at more than 37,000 in 1992 and falling sharply to 22,000 in 1993. This increase is dramatic. Between 1963 and 1976 for example, the number of refugees accepted by Canada had averaged about 3,600 per year (including lows of 600 in 1971 and 1,400 in 1970 and highs of 10,000 in 1968 and 11,000 in 1976). This reflects the increasing numbers of people around the world who have been forced to flee the horrors of famine, civil unrest and oppression. In 1970, the number of refugees worldwide was estimated at 2.5 million. In 1980, this figure had increased to 8 million and by 1993 it was estimated at 18 million.

The numbers of government and privately sponsored refugees has also increased dramatically since 1976. In 1977, 7,300 refugees (government and privately sponsored) were admitted to Canada. Two years later, this number jumped to 27,900 and by 1980 it had grown to a 40,300. Between 1981 and 1986, a period characterised by a major recession, the numbers of sponsored refugee admissions dropped significantly to averages between 15,000 - 20,000 per year. In 1991, this number peaked at 52,300 and has fallen significantly since then.

Since 1976, Canada has experienced two major recessions. The first spanned the late 1970s and the early 1980s and the second took hold in the early 1990s. Immigration numbers reflect this economic picture. For example, in 1974 the total number of immigrants admitted to Canada totalled 218,000; by 1985 this number had fallen to 84,000 (the lowest since 1962). In the late 1980s this number had risen again, reaching

---

4Kelley and Trebilcock 1998:383
5Ibid..347
6Ibid..383
7Ibid..383
250,000 in 1992. Also noteworthy is the change in the ‘complexion’ of Canadian immigration during this period. Prior to 1961, 90% of all immigrants to Canada had been born in Europe and only 3% in Asia and the Middle East. Between 1981 and 1991, 48% of all immigrants were from Asia and the Middle East. Between 1987 and 1990, the numbers of western European immigrants had dropped to 18% of the total. The 1970s and 80s thus witnessed an unprecedented surge in non-white immigrants and refugees admitted to Canada.

Canadians, along with the rest of the western world, responded to these changes with a certain hostility. The growing admission of non-white new immigrants and refugees was not paralleled by growing social acceptance. Over the 1980s, refugees and new immigrants in general were increasingly regarded as a multifaceted ‘threat’ - a numerical threat to be limited in the name of administrative efficiency, fiscal restraint and economic growth; a threat to the ‘integrity’ of the system due to so-called ‘bogus’ claims made by unscrupulous claimants; a threat to national security posed by organized criminals and terrorists; and a criminal threat to the safety of the public.

Racism and xenophobia inflected much of the public and political discussions in this field finding expression and legitimation (of sorts) in the anti-immigration political platforms of, for example, the Reform Party of Canada. One survey after another reported the unsettling findings that Canadians were increasingly hostile towards new immigrants and refugees, and in particular to visible minorities. New immigrants and refugees became the lightening rod and the scapegoat for the lack of jobs in Canada, perceptions of rising crime and the general decline of social order and cohesion.

---

8 Ibid., 382
9 Ibid., 383
ii) Spontaneous Landings, Public Panics and the Fraudulent Claimant

Throughout the 1980s, the dominant preoccupation of the Immigration Department was not with the plight of the soaring numbers of refugees worldwide. Rather what secured their attention was the growing backlog of in-land refugee claims. In popular and political discourse this backlog was then and continues to be primarily attributed to ‘fraudulent’ refugee claims made by ‘bogus’ refugees. For the most part, the ‘bogus’ refugee was seen to be an ‘economic’ migrant who, although not covered by the 1951 Geneva Convention which defined the term refugee, had “...learned that claiming refugee status increased the chance of admission to a more affluent country”\(^\text{11}\). Subsequent to and in part flowing from the Singh decision in 1985,\(^\text{12}\) many refugee claimants were economic migrants. However, given the global conditions noted above, it is unlikely that economic pressures alone were responsible for the huge increase in refugee claims. Indeed a 1986 study conducted by the United Nations High Commissioner for Refugees found that the level of public and political panic centring on abuses of western refugee determination systems by so-called bogus refugees was vastly disproportionate to the actual seriousness of the problem. It reported the number of ‘manifestly unfounded claims’ made in western countries was no more than 10 - 15% of the total claims made.\(^\text{13}\) Nonetheless, the objective of finding better and more efficient ways to dispose of ‘manifestly unfounded’ claims in order to better control, manage and ultimately eliminate the refugee backlog was to be the central focus of and justification for the political and legal actions of the government during this period. In a real sense, the Singh decision provided the government with the political opportunity and legal imperative to restrict access to the system by refugee claimants be they legitimate or ‘bogus’: an objective they pursued with vigilance and determination.

\(^{11}\)Kelley and Trebilcock. 1998:412

\(^{12}\)See discussion in Chapter 2, supra fn. 71, 72

\(^{13}\)

Victor Malarek Haven’s Gate: Canada’s Immigration Fiasco. Toronto: Macmillan. 1987:97
Interestingly, the restrictive, enforcement-oriented exclusionary measures introduced in the late 1980s were usually justified, as will be described, by reference to the still influential humanitarian liberal inclusionary ideals which had underpinned the 1976 Act and its creation of a permanent on-shore refugee determination system. While on the face of it, this increasingly hard-line, exclusionary shift might appear to be inconsistent with inclusionary humanitarian objectives, in actuality it is entirely consistent with liberal humanitarian discourses which privilege and reward ‘deservedness’ and ‘desirability’. Particular conceptions of deservedness and desirability are necessarily, always and already, constituted by their discursive opposition: the figure of the deserving genuine victim exists only in constant and continuous opposition to the figure of the undeserving ‘abusive’ and/or criminal migrant.

In the 1980s, the primary contemporary association of undesirability and undeservedness with criminality and fraud was forged in the public, political and legal realms. Thus, the importance of liberal and humanitarian concerns with deserving victims, in this case the ‘bona fide’, ‘genuine’ Convention refugee, was not in the least inconsistent with or contrary to this emergent discursive shift. Moreover, the rising influence of ‘victim’ discourses which has underpinned ‘law and order’ reforms in the context of the criminal justice system ¹⁴ have bolstered calls for stricter immigration and refugee measures. The argument thus becomes that the undeserving and the undesirable must be more rigorously and effectively excluded precisely in order to ensure that genuine and therefore deserving victims may be appropriately protected. This discursive twist is clearly captured in the comments made in 1987 by then Minister of State for Immigration Gerry Weiner, who defended the hard-line measures taken by the government in the following terms:

We will not tolerate abuse, either within the system of without, because to do so is to deny protection to those who need it and to undermine this country’s long-standing immigration policy itself. We are only too aware of public attitudes and opinions. We will not jeopardize Canada’s

¹⁴ See for example Allison Young’s discussion of the ‘universal victim’ in Imagining Crime 1996: 51-79
openness through complacency or inefficiency...the perception that both our refugee and immigration policies are open to widespread abuse will fuel those [racist and xenophobic] fears. This, we simply will not allow. Refugee policy is for refugees. And within the policy, the person who claims refugee status from Canada, must merit it. This must be clearly understood. Canada's determination system has but one purpose. To distinguish the genuine refugee in need of Canada's help from all other claimants.15

In addition to the increasing hostility toward new immigrants and refugees and to the large and increasingly unmanageable refugee backlog, the adoption of hard-line enforcement-oriented exclusionary legislative reforms in the late 1980s was further facilitated by the eruption of public and political panics around illegal immigrants and refugees. The initial trigger for this panic was squeezed in the spring of 1986 when 155 Tamils from Sri Lanka were found in a boat off the coast of Newfoundland. While the Tamils were relatively compassionately received and quickly granted landing, subsequent news that they had lied about their origins elicited a backlash about 'queue-jumpers' and the adequacy of Canada's control of its borders. In the face of this nasty backlash, Prime Minister Brian Mulroney declared sanctimoniously that "...if we err, we will always err on the side of justice and on the side of compassion."16

Indeed the initial public reaction to the Tamil landing was positively cordial when compared with the veritable explosion of panic, xenophobia, racism and moral outrage which took hold fifteen months later following the 'spontaneous landing' of a boat carrying 173 Sikhs and one Turkish woman on the east coast of Canada. As has been well-described by several authors, the so-called 'Sikh landing' and the public and


political reaction to it provided the government with the justification for enforcement-oriented exclusionary reforms. The rhetoric blossomed exponentially. The government was called upon to respond to this newly constructed 'refugee crisis' which had been propelled by unscrupulous 'queue jumpers', 'bogus refugees' (including primarily the economic migrant but also organized and disorganized, past and modern-day war criminals) and so called 'waves' of naive and witless migrants taken advantage of by international smugglers. A trickle of 174 refugees was quickly likened to a 'wave' - a comparison that inevitably raised fears about the impending 'flood'.

Liberals have an underlying respect for rules - their fairness, their equality of application, their objectivity. Those who respect and obey the rules are commonly morally outraged when others break them. Think of the irritation which is generated during rush hour on a Friday afternoon towards those who shamelessly flout respect for the long-standing practise of taking one's proper place at the end of the line of people waiting for their turn to board a bus. The analogy of 'queue jumping' was quickly deployed in the public discussion of the landed refugee claimant. Moving to the head of a queue before one's turn, whether one is in line for the bus, the cinema or an immigration visa is quickly and easily portrayed as a strong indication of moral undesirability and undeservedness in a liberal society.

The ill will towards the refugee claimant who is seen as a queue jumper is compounded by those who both queue-jump and lie in their efforts to gain admission to Canada. They are seen as showing a complete disregard for the rules, for the liberal sense of fair play. The threatening figure of the immigration 'queue jumper', the 'bogus refugee' the economic migrant (and, to a lesser degree, the criminal - in this case the criminal smuggler) was constructed by and through this public panic. These threatening and undeserving figures were quickly abstracted and generalized and elicited a nationwide panic that was riddled with racism and xenophobia. As put by Julius Grey:

1987 was not a good year for refugees. The arrival of a shipload of illegal

\textsuperscript{17}See for example the discussions in Kelley and Trebilcock 1998; Avery 1995; Mandel 1989; Foster 1998; and Nash (ed) 1988
Sikh claimants, following groups of improbable pretenders from Turkey and Portugal, set off a panic, both in the Canadian government and in the public. Groups of several hundred were likened to "waves" of applicants. Elements of racism, which had never died down, came to the fore...\textsuperscript{18}

Grey refers to the Sikhs as "illegal Sikh claimants". It is an open question as to whether this is an accurate designation. The Singh decision in 1984 had confirmed that all people on Canadian soil enjoy the rights and protections of the Charter and have the right to make a refugee claim and to have that claim heard in accordance with the Charter. The 174 Sikhs who arrived in 1987 sought refugee status. They arrived by boat rather than by plane, and like many refugees were lacking travel documents. It is not clear why they are being deemed "illegal". Presumably it is because it is contrary to the Immigration Act to enter Canada without travel documents - a provision which has been hotly contested in its various applications to refugees who, given that they are fleeing governmental persecution, commonly and understandably do not have the requisite government issued documents.

The labelling of those who land on Canada's shores and who claim refugee status as "queue jumpers" caught on fast. This was the case despite the reality that there is for those hoping to claim refugee status in Canada, very often no "queue" to join, let alone jump. In many cases, the use of this designation which is both defamatory and inflammatory, reveals a lack of knowledge and understanding about the differences between immigrants and refugees:

The "queue-jumping" is explained by the fact that most immigrants have to apply for visas \textit{outside} the country, but most refugee status claimants arrive here and then make the claim. The argument against "queue-jumping would be somewhat stronger if a queue in fact existed. But most immigration applicants are accepted or refused within a relatively short time. There is no list of people waiting for a "place" to become vacant. It is therefore difficult to see over whom refugee claimants are "jumping"\textsuperscript{19}

\textsuperscript{18}Julius Grey "Refugee Status in Canada" in Nash (ed) 1988:306

\textsuperscript{19}Ibid., 306
III) Stemming the 'Flood': The Administrative Crackdown on 'Bogus' Refugees

i) Administrative Measures

After the Singh decision but prior to the Sikh panic, the government had already been seeking ways to counteract the impact of the Singh decision both by trying to deal with the existing refugee backlog and by limiting access to Canada's refugee system. The government felt it necessary to resolve the administrative problem of the backlog without being seen to grant a general amnesty, fearing that such an amnesty would 'open the flood gates' to thousands more 'bogus' claims by economic migrants hoping to be granted another amnesty by a soft government. In fact a partial amnesty was granted in May 1986 which in one fell swoop had reduced the 21,500 refugee applicant backlog figure by 15,000. However it was not long before a second backlog developed and tough talk by Immigration officials about cracking down on abusers became louder.

In February 1987 the government announced several 'administrative measures' purportedly intended to help genuine refugees by cracking down on and deterring abusers. Changes in American immigration law the previous November had made the situation difficult for Central American refugees, in particular those from Guatemala and El Salvador, two countries acknowledged by Canada to be 'refugee producing countries'. The tightening of the U.S. refugee legislation resulted in an increasing number of Central American refugees seeking protection in Canada. Canada responded by sending claimants from the U.S. back across the border to wait for their hearings. This clearly was very threatening to any Guatemalan and El Salvadorian claimants for they might well then be returned by American authorities to their home countries. It can only be interpreted as indicating that the Canadian government was determined deter from making on-shore claims not just 'bogus' refugees but rather the great bulk of potential Latin American claimants.

---


21 Kelley and Trebilcock 1998; Whittaker 1987; Dirks 1995
As well. Canada had already slapped a visa requirement on travellers from both countries (El Salvador in 1978 and Guatemala in 1984) in spite of (or more cynically on account of) the well-documented and widely observed political persecution and terror which characterised them. Malarek reports that at that time, a secret cabinet document revealed that Canada’s response was underpinned by a desire to reduce and deter refugee claims. 22 The government also imposed transit visas on 98 countries to prevent potential claimants from booking flights that include a transit stopover at a Canadian airport with the intention of disembarking in Canada and claiming refugee status.

Visa requirements have long been employed as an immigration control mechanism and their effectiveness is bolstered by the enlisting of airlines in this control function. Visas serve to place a liability on airlines for the return of any passenger who does not have the required documents. The same cabinet document revealed that this measure was aimed primarily at reducing the number of Chileans from disembarking at stopovers in Toronto and Montreal and making refugee claims. 23

A further initiative, also intended to reduce the number of landed claimants, eliminated the ‘B-1’ list which had provided for the issuance of Minister’s permits which would give people from 18 recognized refugee producing countries automatic admission and the right to stay in Canada and work for up to a year. Those claimants already in Canada under these permits were streamed back into the existing refugee determination system despite the growing backlog.

These restrictive administrative measures implemented in February of 1987 shed some light on the way in which the ‘problem’ of the bogus refugee was politically constructed and manipulated. In the name of the controlling ‘abuse’ and deterring bogus refugees, the government justified the adoption of measures to control and reduce the number of all on-shore refugee claims. However, the measures implemented clearly had

22 Malarek 1987:118

23 The Canadian government had already imposed a visa requirement on this legitimate refugee producing country in 1979. Ibid., 118
the undeniable consequence of restricting the access of many deserving and legitimate refugees to Canada’s shores and to its refugee determination system. Many critics attacked the government for its ‘draconian administrative measures’ and charged that the Immigration department had fuelled and then taken advantage of exaggerated public and political concern about ‘abuse’ in its efforts to maintain control over which refugees Canada accepts. As put by the Chair of the Refugee Status Advisory Committee in 1987:

There has been an effort [by senior immigration officials] ...since 1980, to delegitimize the whole process of refugee determination in our country, to characterize this movement as abusive...Their view is that anyone who comes to Canada without prior permission of the Canadian government has abused our process.24

He went on to observe that Immigration officials had deliberately created “a sense of crisis, a sense of being inundated by unscrupulous people who see us as patsies.”25 Winnipeg lawyer David Matas echoed this concern, observing that:

The department isn’t against refugees. What it is opposed to is the notion that somebody can choose Canada rather than Canada choosing the person. So it is trying to set up a system where it decides who enters. and in effect, it is taking advantage of public concern about abuse to get its way.26

ii) Legal Reforms: The Role of Bills C-55 and C-84 (1987)

In May 1987 the government introduced Bill C-55, followed swiftly, two months later, by the ‘Deterrents and Detention Act’, Bill C-84. Together these marked a major mobilization of additional legal mechanisms to augment the additional administrative controls just discussed.

Bill C-55, which was discussed briefly in Chapter One, created the Immigration and Refugee Board, a new two-tiered refugee determination system. This Act, though respecting, to an unprecedented degree, the procedural rights of refugee claimants.

24 Ibid.,121
25 Ibid.,121
26 Quoted in Malarek, Ibid., 122
simultaneously limited access by claimants to the system itself. The central justification which was offered for this limitation was that it would reform the determination system so as to preserve its integrity "...by ensuring the protection of legitimate refugees, while deterring the ‘shameful manipulation’ of false or abusive claims".27

Bill C-55 acted upon the constructed and exaggerated problem of the ‘bogus’ refugee in several ways. It required a person to make a claim at the outset of their immigration inquiry or they would forever be denied eligibility to make a claim later: it significantly expanded the grounds for denying eligibility: it permitted the exclusion of claimants deemed to have come from a ‘safe third country’. the government intending soon to produce a list of such countries: it excluded from making a refugee claim those with a criminal record who had been certified by the Minister as a “danger to the public in Canada”: and it denied a refugee hearing to any who had been granted refugee status in another country or who had already had a refugee claim rejected in Canada. In addition, those found to be eligible still had to receive a decision from the immigration adjudicator and the Refugee Board member that their claim has a ‘credible basis’ (as opposed to the previous standard that excluded only ‘manifestly unfounded’ claims); and, finally it placed contentious limits on access to judicial review of negative determinations of a claimants’ ‘credible basis’ and/or their refugee status. Under Bill C-55 claimants do not have the right to appeal to the Federal Court, rather they are conceded the much more limited right to apply for leave to appeal to the Federal Court of Appeal.

Bill C-84 marks the serious intrusion of the language of deterrence into exclusionary immigration and refugee law and policy. While Bill C-55 aimed mainly to screen out ‘system abusers’ through the tightening up of eligibility and credibility criteria, Bill C-84 which was tabled just over a month after the June 1987 ‘spontaneous Sikh landing’, sought to enhance control of Canada’s borders by providing for tough deterrence oriented penalties for smugglers and others who assist in the ‘transportation’ of illegal immigrants. Bill C-84 increased the government’s power to deal with so-called

spontaneous landings. It increased powers of search and seizure for immigration officials: it imposed fines and jail terms for smugglers and transportation companies: and it even provided authorities with the legal right to turn away boats suspected of carrying refugees before they land.28 Both bills were widely attacked by immigrant and refugee advocates as anti-humanitarian, draconian and dangerous. A representative of Amnesty International, for example, commented that "these measures are designed to keep refugee claimants out of the country as opposed to ensuring that genuine refugee claimants are given protection."

Public and political opinion had been well-primed for an extreme and hard-line reaction before the July landing of the Sikh claimants. The increasing numbers of refugee claimants coupled with the frequently heard representation of these claimants as 'queue-jumpers' and 'bogus' refugees by politicians and immigration officials had already sparked a widely covered national backlash against all new immigrants and refugees. Opinion polls, radio talk shows, letters to newspapers and MPs and a 'rising chorus of so-called ordinary Canadians' expressed increasingly racist and xenophobic sentiments. A banner front page story in the Globe and Mail on March 6, 1987, reported that 80-90% of callers to open-line radio talk shows were strongly against accepting more immigrants and refugees.29

One of the most unsettling aspects of this backlash was its racist nature. The then Minister of Employment and Immigration, Benoit Bouchard, as good as acknowledged this when he identified the increasing hostility to immigrants and refugees as an expression "of a sort of fear" because the new arrivals were "are from such places like

28 A provision which was later dropped in light of the Prime Minister's claim at the time of the Sikhs' arrival that Canada was not "in the business of turning away refugees and we never will under this government." quoted in Kelley and Trebilcock 1998: 417

29 Ibid., 124

30 Malarek. 1987:74
Asia, whereas in the old days they came from Europe."\textsuperscript{31} The Conservative government even called an emergency session of Parliament to "deal with an issue of grave national importance"\textsuperscript{32} and quickly tabled the tough, enforcement and exclusion oriented Bill C-84.

Some have argued that the Immigration Department through willful neglect or intentional mismanagement, contributed to the general panic about system abuse and bogus refugees in order thereby to clear the way for the development of the measures already discussed which limit the numbers of refugees who are able in the first instance to reach Canada and which restricts their access to the determination system once they are here. They point to the Canadian government's failure to impose a visa requirement in a timely fashion on, for example, Portuguese and Turkish travellers to Canada in 1985 which then led to a 'crisis' that could have been easily prevented. When the government did finally require a visa of travellers from Turkey and Portugal, the so-called 'flow' of refugee claimants 'abruptly ceased'.\textsuperscript{33} Prominent lawyer Barbara Jackman argues persuasively that "[T]he Immigration Commission, by becoming incapable of managing the refugee determination system...had used its very mismanagement to justify its wish to close the doors to refugees". With respect to the increased numbers of Portuguese and Turkish claimants. Jackman shrewdly observes that:

In one year, 400 Guatemalans made refugee claims in Canada and 75% of them were recognized as refugees. As a result, visas were introduced to control the flow of refugees into Canada from Guatemala. The government did not hide its purpose. In the case of the Portuguese, the government did not respond until there were 3,000 claimants, until it was publicly known across the country that Portuguese claimants were abusing the process. Only then did they impose visa controls. They used the Portuguese case to justify closing the door to people who abuse the process, yet it was the

\textsuperscript{31}Quoted in Malarek 1987:75

\textsuperscript{32}Kelley and Trebilcock 1998:417

\textsuperscript{33}Ibid.,415
government that let the situation develop.\textsuperscript{34}

As noted in the opening paragraphs of this chapter, Canada had never been keen to become a country of first asylum and has been determined to remain to as full an extent as possible a country of resettlement for refugees it chooses. This continues to be a major influence on government policy and practice. In 1983, senior Immigration official Raphael Girard, who would later be the principal drafter of Bills C-55 and C-84, spelled out this position in a departmental discussion paper: "...it is not desirable to have a resettlement program straddling two main themes, active off shore selection and the use of asylum as a pro-active program." He argued that the main emphasis of Canada's refugee program should continue to be off-shore selection. His arguments were that off-shore selection favoured those who could 'best benefit' as opposed to the self-selected, was responsive to domestic concerns and priorities, could be easier 'managed and controlled', was amenable to foreign policy objectives, and did not entail the problems associated with removing unsuccessful claimants.\textsuperscript{35}

The public reaction which developed in the late 1980s around the threat posed by the so-called 'bogus' refugees, provided the official justification for the restrictive and exclusionary dimensions of Bill C-55 and Bill C-84. However, as described above these measures were not designed to facilitate access for genuine refugees, nor to exclude only those who could not credibly claim to be refugees as defined in the International Convention. They sought to limit the numbers of all on shore refugee claimants.

This analysis can now be advanced a further stage. It is true that these measures were usually justified as a protection for genuine refugees and that this illustrates the continued discursive importance of humanitarian liberal ideals. Nevertheless, and more significantly, this discursive power of the deserving 'genuine' refugee was maintained

\textsuperscript{34}Nash (ed.)1988:323

and fortified only in contradistinction to the increasing influence of the socially, politically and legally constructed figure of the undeserving and unscrupulous figure of the fraudulent claimant - the economic migrant, the 'queue-jumper', the 'bogus' refugee. Weiner, the Minister of Immigration illustrated this important observation when he declared:

The thrust of Bill C-55 was to streamline Canada’s refugee-determination system so as to maintain ‘the integrity of our refugee determination system’ by ensuring the protection of legitimate refugees while deterring the ‘shameful manipulation’ of false or abusive claims.” 36

Indeed, while this legislation was explicitly defended *in the name of* genuine refugees, absolutely nothing in the legislation sought to actually facilitate the admission of genuine refugees. Rather it was simply assumed that humanitarian ends would be furthered by cracking down on fraudulent claimants. All this parallels and is bolstered by the rising influence of victim discourses which have underpinned calls for hard-line law and order reforms to the criminal justice system. In each arena the undeserving and undesirable must be effectively excluded precisely in order to ensure that the genuinely deserving victims may be adequately protected, be they law abiding citizens of our own society or the victims of state sanctioned persecution in other countries.

IV) Expanding the Threat in the 1990s: ‘Bogus’ Refugee Meets the Terrorist and the Serious Criminal

Thus far, this chapter has traced the emergence of the ‘bogus refugee’ as a discursive regulatory figure. Once named and identified, a problem may be acted upon: it becomes an object of regulation. The bogus refugee elicited a backlash against immigrants and refugees that was widespread, self-righteous, and racist. This backlash was further heightened by a discursive link which was made between ‘bogus refugees’ and terrorists and/or criminals. During the 1980s, a major threat to national security was

presented and represented by the figure of the international terrorist. Whittaker identifies this development well:

International terrorism is the great frisson of the decade. not unlike the great fear of international Communism that gripped the Western world, and especially the United States, in the late 1940s and early 1950s. Terrorism presents, of course, a real threat to security and public safety. But as in the earlier case, there is also much exaggeration about it, much panic, and much talk of extreme measures. There is no shortage of self-proclaimed 'experts' on terrorism prepared to sell their nostrums to a fearful public. And there is no shortage of voices counselling the abandonment of liberal freedoms and the need for stern repressive measures.

During 1986, the perceived threat of terrorism provided the focus for several governmental initiatives including the striking of a special governmental committee on terrorism and the holding of a conference on terrorism in Quebec attended by military, security and police officials. In both arenas terrorism was regarded as a serious and growing threat to Canada, and in both arenas the refugee determination system was identified as a 'security' problem. The special committee gave particular attention to 'terrorist prone' ethnic groups and the conference attended to similar preoccupations. Not surprisingly, Sikhs were identified in both contexts as a particular terrorist threat.

The threat posed by international terrorists which had emerged over the 1970s and 1980s became the discursive replacement of the threat previously posed by domestic subversives during the height of the Cold War. Thus, while the numerical threat posed by 'bogus' refugees provided the primary and immediate justification for increasingly exclusionary and enforcement oriented measures, the fears sparked by the image of 'waves', 'floods' or 'tides' of bogus or fraudulent claimants were heightened by a discursive linkage made with the threatening figures of the terrorist and the 'serious' criminal immigrant or refugee, posing a fresh and additional risk to the state and the

37 Whittaker 1987:300

38 Ibid., 300
nation’s public.

Immediately upon the arrival of the Sikhs in June 1987, an ‘orchestrated hysteria’ was created by the Minister of Immigration, the Immigration department and the Prime Minister to incite public opinion. They made many inflammatory statements about the Sikhs including comments implying that they were serious security threats and terrorists. As observed by Jackman:

They took advantage of these events by not reacting in a responsible manner, particularly with the Sikhs. Instead, they lead people to believe the Sikhs were dangerous security threats and terrorists. Thus they pandered to public xenophobic reaction and produced a rise of racism.

Less explicitly but no less significantly, the Tamil landing of 1986 was also linked with international criminal and subversive organizations. This linkage of the ‘bogus’ refugee and the criminal or terrorist threat is clearly made in the following defence of Bill C-55 and Bill C-84 by Jim Hawkes, then MP for Calgary West:

...We need quick decisions if we are going to have a system of law in which protection is offered to legitimate Convention refugees and abusers can be removed quickly and efficiently. In a world of increasing international terrorism their has to be a means of determining and detaining people.

Bill C-84, as already explained, sought to deal with so-called ‘spontaneous landings’ by acting primarily upon smugglers and transportation companies. However, it also provided some fairly serious ‘security-related’ changes. First, Bill C-84 excluded anyone certified by the Minister and the Solicitor General as a danger to the public due to criminal convictions, this certification not being reviewable. Secondly Bill C-84

39See Jackman and Heap in Nash, 1987
40Jackman in Nash 1987:322-323
41Ibid..323
42Avery 1995:221
43Hawkes in Nash (ed.)1987:255
sanctioned the detention of undocumented persons upon their arrival in Canada until such a time as their identities could be confirmed, a measure approximating indefinite and therefore unconstitutional detention. This arbitrary detention, defended on security grounds, was likely to affect most negatively refugee claimants who often do not have the required documentation. 44

The Bill also provides for the detention at the border of anyone thought by the authorities to pose a security threat. Both groups may be detained for up to 28 days without review. Persons thought to be a security threat are likely to be detained for even longer for they become the subject of a security review procedure entailing that they appear before a federal court judge. While anyone who is the subject of such a review may respond to the judge's summary of the case, they have no right either to hear or confront the witnesses or to challenge the evidence presented against them. If the security threat is judged to be real, the Bill provided for the exclusion and immediate deportation of the person in question. These provisions constitute an application of article 32 of the International Convention Relating to the Status of Refugees which allows for the refoulement of refugees (the return of people to a place where they would face persecution under the Convention definition) only in those cases where they are duly found to represent either a 'security' or a 'serious criminality' threat to the nation and/or public.

What is remarkable about the two Acts of 1987 and centrally important for the argument of this chapter is that they act upon, and arguably thereby link, three distinct discursive categories of undeserving and undesirable non-citizens: frauds, terrorists and criminals. Each represents a significant threat to the integrity of the state's administrative systems, to commerce, to government and to the safety of the public respectively. This linkage became increasingly dominant in the early 1990s.

44 see Jackman in Nash 1987:322-323
V) In the Name of the ‘Truly’ Deserving: Bill C-86, The “Deterrence and Detention Act” (1992)

The discursive linkage of criminality and fraudulent abuse became front and centre with the next major piece of immigration legislation, Bill C-86, the *Deterrence and Detention Act* in 1992. Danger and criminality discourses, already well-established in the context of domestic criminal justice law and policy were mobilized with a vengeance in the context of immigration and refugee law and policy. Thus in addition to emphasizing the need to protect Canadians against those outsiders who “abuse the system” through fraudulent claims, Bill C-86 ‘responded’ in a major way to the growing ‘problem’ of the criminal immigrant and/or refugee.

Refugees had already been increasingly criminalized through the mobilization of fraud discourses. Since the 1976 Act, the regulatory figure of the criminal immigrant encompassed an ever-widening range of threatening criminal ‘types’. These have included: international terrorists seeking to found new bases for terrorist organizations; war criminals, either the ‘modern day’ or the traditional (Nazi) variety; ‘serious’ criminals with ‘serious’ criminal histories; and the relatively new category of organized criminals.\(^4\) The exclusionary conception of ‘criminality’ as grounds for denying immigration or refugee applicants has expanded. It is now understood and acted upon as a threat to the security of Canada and of Canadians, thus justifying increasingly repressive political and legal responses. Similarly, the notions of ‘system abuse’ and the lack of ‘control’ over the system which allows such abuse to flourish, which were previously primarily linked with the numerical and moral threat posed by the ‘bogus’ refugee, has been discursively extended and linked with the threat posed by terrorists and other dangerous criminals:

An immigration program that is not properly controlled is vulnerable to abuses by criminals, terrorists and others who might jeopardize the safety and well-being of Canadians. In recent years we have seen the

\(^4\) It is analytically relevant that the mandate of CSIS was later expanded to encompass organized crime and criminals as a threat to national security.
development of more organized, highly professional criminal networks intent on circumventing international and national laws. As the volumes of people seeking to enter Canada increase, vigilance is needed to ensure that Canadian society is protected from those who are not welcome in our country and who are intent on breaking its laws. 46

Bill C-86 was represented as a new 'managerial' approach to immigration necessary in the face of a global context characterized by "...growing, unpredictable, and large scale movements of people from one country to another" 47. The policy paper just quoted set the stage for the Bill's restrictive and enforcement-oriented provisions by painting a global picture which would likely evoke fear and insecurity in Canadians. The image of 'volumes' of internationally displaced, dispossessed and needy foreigners clamouring at 'our' borders is conjured up as a potential threat to the Canadian state, its administrative systems and its public. Sheer numbers of global migrants are emphasized in dramatic and ominous fashion: "[S]ome estimates suggest that today, as many as eighty million people-more than three times the entire population of Canada-are moving from one country to another at any given time." 48

Specific reference is also made to the fact that although the numbers of refugee claimants has increased exponentially, "the proportion of claimants who are found to be convention refugees is falling." 49 An observation made in the context of a hard-line piece of British refugee legislation in 1993, may be applied to the Canadian scene. A "culture of disbelief" about the legitimacy of claims to asylum has been effectively created and "...that is the reason for the dramatic decline in the proportion of asylum seekers granted refugee status or exceptional leave to maintain. Those figures are then used to justify the


47 Ibid..3

48 Ibid..4

49 Ibid..4
belief that most asylum seekers are ‘bogus’."

As documented by Jakubowski, numerous polls and surveys carried out at this time of recession provide evidence that Canadians were ‘...feeling far more vulnerable economically. [were]...developing a growing sense of mistrust and intolerance towards outsiders. and...[were] increasingly disillusioned and discontented with the government.’ Moreover, it was commonly reported that many Canadians were concerned about the increase in immigration during a period of economic recession and in particular about the increase in non-white immigration and refugee admissions.

In this less welcoming environment, the Canadian government, rather than attempting to diffuse some of these hostilities and fears, mobilized and exploited them for their own purposes. For example, the Minister of Immigration, Bernard Valcourt, publicized the findings of two 1992 government sponsored studies which highlighted the xenophobia felt by many Canadians towards immigrants and refugees. One of these studies reported that one third of its respondents had agreed that it was important to ‘...keep out people who are different from most Canadians’ and that nearly half were ‘really worried that they might become a minority if immigration is unchecked.’ Valcourt’s tactics are summed up nicely in the following journalistic observation:

Valcourt has used some strong language in comments defending the bill. He’s evoked images of millions of refugees from Third World countries fleeing poverty and turmoil, implying that they may soon be clamouring at our door. He’s warned that the government needs more power to keep criminals and terrorists out of the country. He’s told the story - several times - of a Montreal refugee claimant who filed 14 fake claims in order to cheat welfare.

---


51 Jakubowski. Immigration and the Legalization of Racism 1997:66

52 Ibid. 66-70

53 quoted in Jakubowski 1997:70

54 The Citizen Valley. "Complicated, Confusing and Controversial" November 23, 1992
Immigrants and refugees were generally represented as a threat which needed to be contained and controlled; "[I]n trying to sell Bill C-86 to Canadians, Valcourt play[ed] on their insecurities by socially constructing the immigrant or refugee as a potential system abuser of whom we should be wary." In Valcourt's own words, "Canadians are compassionate and humane. but we don't want to be taken for a ride."

Bill C-86 thus promised to reduce significantly the opportunities for 'system-abusers' and 'criminals'; by introducing this new, tighter and more efficient managerial immigration regime. The predominant policy objective was to protect society. And once again, one of the main justifications for this 'toughening up' of the system was putatively humanitarian: it was necessary to crack down on the undeserving and undesirable in the name of the genuinely deserving bona fide, convention refugee: "The proposed changes...provide for a more streamlined refugee determination system. ensuring that we can help those who truly need refuge..."

Bill C-86 radically extended the criminality-based exclusionary grounds for inadmissibility. With this Bill, immigration law, policy and practices are represented and justified as the thin blue line protecting Canadians from the wide ranging criminal threats which lurk just outside of our borders. Canadian immigration law and policy have emerged in the 1990s as a critical and necessary instrument of crime control to an extent

---

55 Jakubowski 1997:71


57 Immigration Canada, "Explaining the New Immigration Bill" 1992:preface (my emphasis)

58 Bill C-86 also extended the grounds for medical inadmissibility. again under the general rubric of 'protecting' Canadians from potential sources of danger. While not directly relevant to the current analysis, it is noteworthy that the image of the 'diseased' immigrant (in the 1990s tuberculosis is the dominant looming threat carried by the 'volumes' of the poor and needy masses seeking access to Canada) is once again gaining discursive prominence in exclusionary immigration law and policy. It is foreseeable that this developing 'problem' will increasingly provide the justification for more restrictive exclusionary measures.
Bill C-86 sought to fill perceived gaps in the existing legislation with respect to the inadmissibility of criminals 'and others who threaten the security of Canada' (sections 19 and 27 of the Act). Specifically it aimed to exclude 'organized criminals' previously unaddressed in the Act. As explained in a governmental guide to the new legislation, existing legislation was inadequate because it did "...not directly enable immigration officials to refuse admission to persons who may have no criminal convictions, but who are nevertheless involved in organized crime or other criminal activity, according to foreign police reports or intelligence sources."

In response, Bill C-86 expanded the definition of criminal inadmissibility to include where there are 'reasonable grounds to believe' the person was or is a member of an organization 'that there are reasonable grounds to believe' is or was engaged in activity "that is part of a pattern of criminal activity planned and organized by a number of persons acting in concert in furtherance of the commission of any offence that may be punishable...by way of indictment." (s.19 (c.2) of the Act) Moreover, it also allowed for the exclusion of persons where there are 'reasonable grounds to believe' that they have committed an 'act or omission' outside Canada that would constitute an offence inside Canada. (s.19 (c.1)(ii)) 

Bill C-86 also addressed the perceived weakness of the existing legislation to exclude terrorists. As observed in the guide to the legislation,

---

59 With respect to its inclusionary preoccupations, Bill C-86 was primarily supported by economic arguments. It sought to enhance the responsiveness of immigration law and policy to the 'economic and labour force needs of Canada'. For example, as put by the government, investors must be accepted "without limits" for their contributions to the economy. Bill C-86 also made much of the government's humanitarian objective of family unification, arguably hoping to tap into the increased influence of fundamentally conservative and traditional 'family values' evident in the broader social and political context.

60 "Explaining the Immigration Bill" Section 2: "Protecting Society", 1992:3
Current provisions do not explicitly refer to terrorists, although reference is made to espionage, subversion, and group activity involving violence against persons. Amendments would define terrorism as activities directed at, or in support of, acts of serious violence against persons or property for political ends.\footnote{Ibid., 4}

Bill C-86 authorized immigration officials to bar entry and/or deport individuals who have engaged in or who, upon ‘reasonable grounds’ are believed may engage in espionage, subversion and/or terrorism as well as those who are members of an organization which has, or is reasonably believed, may engage in espionage, subversion, terrorism and/or “other acts of violence that would or might endanger the lives or safety of persons in Canada” (ss. 199(e)(f)(g)). The bill also included a new ‘catch-all’ ground for exclusion on security grounds. S.19 (k) sanctions the exclusion of persons “who constitute a danger to the security of Canada” but who are not covered by existing grounds. Even permanent residents may be deported under these provisions.\footnote{62}

The Bill also introduced s.19(1) of the Act to facilitate the exclusion of people who are or were “senior members of or senior officials in the service of a government that is or was engaged in terrorism, systematic or gross human rights violations or war crimes or crimes against humanity...” The section goes on to provide a ‘general’ guide to the definition of ‘senior members’ or ‘senior officials’, but makes clear that the list of positions and/or functions is not exhaustive. This list includes: a) heads of state or government; b) members of the cabinet or governing council; c) senior advisors to either heads of state, cabinet or governing council members; d) senior members of the public service: e) senior members of the military, intelligence and/or internal security apparatus: f) ambassadors and senior diplomatic officials; and g) members of the judiciary.

\footnote{Ibid., 4}

\footnote{62}

A comment made in the Parliamentary Committee hearings on the Bill nicely sums up one of the central criticisms made about these amendments. Alan Borovoy of the Canadian Civil Liberties Association described them as sanctioning “deportation by clairvoyance” \citep[quoted in][]{Kelley2004:431}
In addition to extending the criminality provisions of the Act, Bill C-86 toughened up the provisions and penalties relating to the transportation of illegal migrants to Canada (ss. 85.86.89.1. 91.1. 92. 92.1. 93. 97.1) and the smuggling of illegal immigrants into Canada (ss.94.1. 94.2. 102.01(1)). For the present purposes, one final set of amendments needs to be mentioned, those pertaining to border ‘controls’ (s.110). These provisions gave immigration officers the discretion to fingerprint and photograph refugee claimants and expanded their search and seizure powers. The latter amendment sanctioned the search of person, luggage and/or vehicle when there are ‘reasonable grounds to believe’ the person has hidden their identity documents. The same authorization is given to officers to search ‘...persons seeking to come to Canada who are believed on reasonable grounds to be smugglers, document couriers, and others involved in the illegal entry of persons...’

Bill C-86 thus ushered in a new discursive era in the governance of immigration and refugee law and policy. It broadly defined and acted upon the emergent threat of the criminal immigrant to a degree unparalleled in foregoing immigration legislation and it similarly enhanced the coercive policing powers of immigration officers to enforce the Act.

As was true for Bills C-55 and C-84, Bill C-86 was greeted with extensive and well-organized criticism from a wide range of legal and non-governmental groups. Much of that criticism focussed on the ‘draconian’ security and criminality provisions of the Act. However, many of the changes made to the existing refugee determination system were also sharply criticised. Of particular concern to non-governmental representatives was the revival of the safe-third country provision which had been included in Bill C-55 but which had never been acted upon by the government. This provision disallowed refugee claims made by individuals if they had travelled to Canada

---


either directly or indirectly from a country prescribed by the governor in council as “safe”. A “safe” third country is one which complies with article 33 of the Refugee Convention which prohibits *refoulement*. Objections to this provision were extensive and have been discussed in some detail elsewhere. For the present purposes it is significant that this provision, if and when it is operationalised, “would systematically exclude those refugee populations from the developing world (most of who are visibly different) who cannot get direct flights out of unsafe countries.”

The racist impact of this Bill was widely criticized: “Critics have called it mean-spirited, sinful and racist.” One third of all refugee claims are made along the border with the U.S.. Were the U.S. to be prescribed (despite its poor and highly politically partisan record of refugee determination), the number of refugees who could make their claim in Canada would be dramatically reduced. As reported in an Ottawa daily newspaper. “[T]hat could mean Canada would turn away up to 75% of the 30,000 refugee claimants who arrive at our border each year because they travel via the United States or Europe.”

The safe third country provision, along with most of the other exclusionary amendments of Bill C-86, evidences once again the Canadian government’s determination to limit as fully as possible the numbers of refugees (genuine or otherwise) who arrive on the shores of Canada and seek acceptance as a Convention refugee.

Canada’s ‘humanitarian’ commitment is thus at best highly selective and inconsistent. As evidenced by Bill C-86 and as expressed by many critics, the government, was more concerned to keep refugees out of Canada than it was with protecting them. To quote Canadian immigration lawyer David Matas:

...the bill itself does not appreciate the refugee issue from the angle of refugee protection...Really, immigration management and refugee

65 see for example Jakubowski 1997:81-89

66 Ibid., 1997:85

67 *The Citizen Valley* “Complicated, Confusing and Controversial” Nov.23. 1992

68 Ibid.
protection are two different things and require two different approaches. Immigrants are people who are coming to Canada to settle permanently. Refugees are people being forced to flee a situation of danger, often on a moment's notice. Management suggests that Canada is choosing its immigrants in a planned way. Yet refugees in fact, in law and by the terms of Canada's international obligations are self-selected. Managing immigration really means, in a refugee protection context, denying protection to people who otherwise might be allowed and entitled to come as refugees. That is indeed the effect of many of the provisions in the bill. They make it more difficult for refugees to seek protection or even to arrive in Canada to make a claim for protection. 69

Finally, the new legislation eliminated the first stage credibility and eligibility hearing introduced by Bill C-55 to weed out 'system abusers' at an early stage in the process. 70 Bill C-86 instead gave Senior Immigration Officers the power to rule on refugee eligibility, again an administrative decision replacing a quasi-judicial one. The government sought to minimize the significance of this expansion of the discretionary power of SIOs. It asserted that the eligibility decisions to be made by SIOs will made on the basis of a checklist of 'facts' (rather than on 'matters of judgement') and will therefore be entirely non-discretionary. 71

Thus it can be seen that the governmental logic of national security (encompassing more and more varieties of criminal activity) and the corollary notion of protection of the

69 Quoted in Jakubowski 1997:79

70 Interestingly, the number of clearly unfounded 'bogus' claims was not readily apparent to Immigration officers who had referred a substantial 94% of claims to the CRDD for a full hearing Ibid.,81

71 This checklist denies claimants access to the refugee determination system on one of five grounds: a) prior recognition of refugee status in another country; b) coming to Canada directly from or through a prescribed 'safe third country'; c) repeat claims; d) have already been determined in Canada or by a visa officer abroad to be a Convention refugee; or e) "undesirable persons- criminal and security risks" (Immigration Canada. "Explaining the New Immigration Bill" Section 3: The New Refugee Determination System, subsection 1: Access to the Refugee Determination System,1992:2)
public provided the dominant rationale for the exclusionary provisions of Bill C-86. The ‘problem’ of the ‘bogus refugee’ had been legally constructed by and acted upon through Bills C-55 and C-84. It re-emerged with Bill C-86 as the more general problem of ‘system abuse’ which was attributed in large part to a lack of adequate control over the immigration system. This lack of control was in turn linked with the more dominant threat of the 1990s that posed by the criminal refugee or immigrant broadly defined. Bill C-86 legally constructs and acts upon these two ‘problems’. system abuse and criminality, by sanctioning ever more repressive and exclusionary means: “With Bill C-86 we introduced a wide range of tools to deal with those who try to abuse our immigration program or violate our laws.”

VI) The Creation of the Canadian ‘Public Security Portfolio’ in 1993

In one of the more honest and transparent political decisions of this period, in 1993 the Conservative government, on the heels of Bill C-86, created a new Public Security Portfolio. This brought together Immigration, the RCMP, CSIS and the Canada Parole Board into a newly created government department, the Department of Public Security. The first Minister of Public Security, Doug Lewis, explained the logic behind move by reference to the “...global community [which] is awash with people seeking new homes” (again the water metaphor conjuring up a disastrous flood) and the consequent the need to “reinforce and defend the integrity of our generous and valuable immigration policy”: to “give extra weight to the issue of enforcement”; to have “more efficient control over the security of our borders”; and to better protect Canadians.

By consolidating management of our border activities and our immigration enforcement activities, I am convinced that we can exercise more effective control over entry to Canada; ensure that we better protect all Canadians; and reduce abuse of Canada’s generous immigration and refugee programs.”


Lewis. “Notes for an Address” 1993
While the discourses of fraud and criminality are employed in this political justification, primary emphasis is given to populist reasons for the switch. Lewis quite freely admitted that of the 220,000 newcomers who arrived in 1991, “only a very tiny percentage of those arrivals were mired in controversy”. He even lamented that many Canadians don’t share this view and acknowledged the powerful and deleterious effect of exaggerated, sensational and inflammatory coverage of “...the isolated case of people who abuse our generosity”. What is curious and indeed telling, is the determinative value placed by Lewis on public support with respect to the changes at hand. It is the need to restore ‘public confidence and support’ that seems to underpin this move, rather than any real or substantiated threat posed to either the system or the public: the choice, according to Lewis is straightforward: “We can either reinforce and defend the integrity of our generous and valuable immigration policy or we can watch public confidence and support for the policy collapse.” 74

This move was not a subtle one. Critics charged that conservative Prime Minister Kim Campbell was fanning the populist flames against new immigrants and refugees by “baiting Canadians fear that immigrants commit a disproportionate amount of crime. and by...linking the solution to the perceived crime problem with cleaning up the immigration and refugee system.” 75 As succinctly stated by Shyla Dutt of the Asian Canadian Caucus, the government had succeeded in “re-focussing public debate away from immigration as an economic and social issue, and [toward immigration] as a security and public safety issue.” 76

Responding to these complaints, Lewis likened this administrative and organizational decision to that which gave police officers the authority to enforce

74Ibid.
75Titch Dharamsi, The Metro World "‘Fighting for Inclusion’: Fighting for the Anti-Immigrant Vote” October 1993: 7.29
76Ibid., 7
There are those who claim that, by placing elements of the immigration program in the Public Security Portfolio, we are ‘tarring’ all immigrants and refugees as criminals or as ‘threats’ to public security. I am honestly puzzled by the logic of this argument. Police officers are responsible for enforcing highway traffic regulations. Does this imply that all drivers are criminals? Of course not. And it doesn’t mean that the police cannot protect the vulnerable of our society and promote the good of society...It is surely just as important [as enforcing traffic laws] to maintain the integrity of our borders and to enforce Canada’s Immigration Act in the most effective way possible.”

VII) Conclusion

Just as Bills C-55 and C-84 acted upon (and reproduced) the ‘bogus refugee’ as the archetype of moral and legal ‘undeservedness’, Bill C-86 acted upon the ‘criminal immigrant or refugee’ as the archetype of moral and legal undesirability. The victim of the fraudulent character of so-called bogus refugees has been constructed as the state, specifically state administrative systems and ‘Canadian generosity’. It is similarly the case that different forms of criminality have been incrementally but steadily redefined as posing a risk/danger to the security of the Canadian state and of the public. Once so viewed, criminal and security risks may then be ‘legitimately’ acted upon through increasingly repressive and enforcement oriented sovereign power.

By 1992, the criminal immigrant or refugee was produced as the archetype of undesirability. It now stands alongside the archetype of undeservedness, the ‘bogus refugee’. By 1992, the figure of the ‘bogus refugee’ was mobilized more as evidence of a poorly managed system open to abuse by criminal and security risks rather than as a dominant problem in and of itself. In the following chapter, the discursive redefinition of refugees (from deserving victims to undeserving frauds/criminals) will be examined in more detail as will the material impact of these developments on the lives of particular groups of new immigrants and refugees in Canada.

77 Lewis. “Notes for an Address” 1993
Chapter Five  
From Deserving Victims to Undeserving Criminals:  
Redefining and Regulating Refugees in the 1990s

Security is the great commodity of our time...Government is virtually defined by the problem of managing the security of the population...For those outside the limited access environments in which these new security mechanisms proliferate, the hard edge of traditional social control remains the predominantly experienced means of security. But this apparatus of security is decidedly ambiguous. Its aim is both to secure the underclass and to secure others against it.  

1) Introduction  
On the heels of yet another recession in 1993, and in the face of the massive displacement of people globally, it is not surprising that the issues of welfare and immigration would be considered in relation to each other. But what is perhaps less straightforward is the way in which the discursive figures of the ‘welfare cheat’ and the ‘bogus refugee’, constructed by and through panics and crackdowns in both the immigration and social services contexts, merged and were mobilized in distinctly punitive and coercive ways, ways which, moreover, particularly and adversely affected visible minorities. Alongside the increasingly powerful construction of the ‘criminal immigrant and/or refugee’ (in spite of consistent statistical evidence that immigrants and/or refugees are not over represented in the prison system, and despite a steady decline in national crime rates), over the late 1980s and early 1990s, the regulatory figures of the

1Jonathan Simon Poor Discipline Chicago: University of Chicago Press. 1993:258 2
As observed by sociologist Morton Weinfield in dialogue with Daniel Stoffman. “Many people do make the linkages between immigrants and crime. In fact, if we look at the evidence...you find, for example, in the federal prisons that the foreign born are under represented as a proportion of the inmate population. So the whole notion that immigrants are heavily involved in crime just doesn’t withstand scrutiny.” (“How Many Immigrants Should Canada Admit” Globe and Mail, March 4, 1994) For a comprehensive look at the relationship between immigrants and crime see “Immigration and Crime” Allan Borowski and Derrick Thomas in Adelman et al (eds.) Immigration and Refugee Policy: Australia and Canada Compared, Toronto: University of Toronto Press. 1994; “Race. Ethnicity
welfare cheat and bogus refugee emerged, both within and without 'our gates'. Social and political anxieties about the threat posed by criminal immigrants and refugees merged with the public and political panic about the 'welfare cheat' and the 'bogus refugee' to produce a powerful backlash against certain visible groups of new immigrants and refugees in the 1990s. On one level, these discursive developments can be understood as the incremental but steady redefinition of the 'claimant' of the benefits of administrative justice, from that of a deserving victim 'at risk' to an undeserving, risky 'offender' (whether represented in terms of 'fraud' or 'criminality' or a rather explosive blend of both).

This shift in turn reflects the broader shift in governmental rule, from humanitarian welfarist liberal to neo-liberal. This more general shift in the rationality of governmental rule has entailed the emergence of different governing logics. In the context of exclusionary Canadian immigration law and policy, in the 1960s and early 70s, the inclusionary logics of human rights and liberal legalism challenged and supplanted the previously dominant logics of immorality and national 'purity'. Since then, the logic of national security has been incrementally expanded to include more and more forms of 'criminality' which were themselves redefined as posing a threat to national security and/or protection of the public.

While the logic of national security was certainly firmly in place when the 1976 Act was adopted, it encompassed the more traditional Cold War concerns with threats to the Canadian political state posed by political subversives, terrorists and those engaging in espionage. The 1976 Act articulated the new 'security' concern of organized crime; however this concern was still in its early stages. Since 1976, the logic of 'security' has underpinned and justified the steady intensification, specification and in some cases redefinition of a wide range of new threats, those posed by organized criminals, international terrorists, domestic political terrorists, war criminals (traditional and modern...
day), persons connected with regimes who have perpetrated wide scale human rights abuses, and dangerous criminal offenders. Moreover, the logic of ‘protecting the public’ has also become increasingly dominant and has been discursively linked with the notion of ‘national security’. Protecting national security and protecting the public have emerged over the last three decades as the guiding rationales of exclusionary Canadian immigration law and policy. The rise in dominance of these rationales has entailed the increased mobilization of the constitutive discourses of criminality, victimhood and risk/danger. As discussed above, the logic of ‘system integrity’ and its corollary, the constitutive discourses of abuse and fraud, also emerged during this period in a manner most consistent with the general decline of welfarist humanitarian liberalism and the rise of a more punitive neo-liberal mode of governance.

This chapter begins with a brief discussion of the rise of neo-liberalism as a rationality of rule and the criminalization of certain kinds of fraud. It traces the emergence of the problem of the ‘welfare cheat’ in Ontario in the early 1990s. It then examines the ways in which the discourses of fraud and abuse converged with criminality discourses in the construction and regulation of the fraudulent criminal immigrant. This discursive redefinition of deserving refugee ‘victims’ to undeserving abusers, frauds and/or criminals will then be considered in terms of the direct and indirect racist and coercive consequences that this shift entailed and indeed facilitated for Somali refugees living in Toronto who became the specific subjects of this redefinition in public and political realms and who consequently were the objects of increased coercive control and legal regulation.
II) The Rise of Neo-Liberalism as a Rationality of Rule

Look upon fraud as a crime greater than theft... for they allege, that care and vigilance, with a very common understanding, may preserve a man's goods from thieves but honesty hath no fence against superior cunning... where fraud is permitted or connived at, or hath no law to punish it, the honest dealer is always undone, and the knave gets the advantage.  

The linkage and increasing dominance of fraud and criminality discourses over the late 1980s and early 1990s is not limited to the field of immigration and refugee governance. Indeed, in the early 1990s, provincial governments were increasingly preoccupied with fraudulence in the context of social welfare provision. Alongside the construction of the ‘bogus refugee’ emerged the regulatory figure of the ‘welfare cheat’. In the 1990s, the figures of the ‘bogus refugee’, the ‘welfare cheat’ and the ‘criminal immigrant or refugee’ came to be closely linked in popular and political discourse.

Like the figure of the fraudulent refugee claimant, the fraudulent welfare claimant evokes considerable moral indignation and condemnation. Taking advantage of Canadian generosity, cheating the system, getting something for nothing, both the bogus refugee and the welfare cheat easily offend many of the essential qualities of the desirable liberal citizen: autonomy, (economic) independence, acquisitiveness, merit, dessert, accountability and rule-governed. The spectre of fraud is mobilized as a technology of liberal governance: it is employed in the continuous and varied project of producing and reproducing governable - self-regulating - citizens.

The logic of fraud enjoys a heightened discursive influence in most areas of public policy under a rationality of rule that has shifted from a humanitarian welfarist orientation to a neo-liberal one. Neo-liberalism grew in influence as the post war boom came to a close in the 1970s and high inflation, unemployment and recession marked the next few decades. Since the 1970s, both the economy and public discourse have changed

---

3 Gulliver's Travels, Swift, 1726 quoted on the 1999 Web Page of the Toronto Police Service Fraud Squad
dramatically. Neo-liberalism seeks to reassert pre-Keynesian classical economic theory. The classical liberal commitment to a minimal state has been revived. neo-liberalism seeks to shrink the size of the state and restore the primacy of market forces. The 'welfare state' must be retrenched if not completely dismantled as it is wasteful, excessive, a hindrance to wealth creation and to the state's ability to compete economically in international markets.  

This neo-liberal rationality rests on a blend of economic, political and moral arguments. The claim is that the welfare state has reduced freedom by enforcing financial redistribution from taxpayers to welfare recipients. It has eroded the work ethic and made recipients dependent. It has created a huge public sector which has a vested interest in its perpetuation. That situation is viewed as untenable in face of massive economic global restructuring: "market forces must be liberated, by downsizing the state and by emasculating social welfare policies, so that the state can meet the challenges posed by global restructuring." Neo-liberalism holds that economies must adjust to the changing opportunities that are available to them in the global economy. These adjustments certainly unsettle labour markets, rendering employment more uncertain and insecure. However, these features of the labour market are legitimized as 'unavoidable', 'normal' even 'desirable'. A belief in the superiority of the market as a mechanism for organizing society underlies the economic propositions of neo-liberal doctrine. State intervention is an impediment to the smooth functioning of the market:

The logic of the doctrine of market superiority requires the state to remove

---


5 Ibid. See in particular: Philip Resnick "Neo-conservatism and Beyond"; Robert Mullaly "Social Welfare and the New Right: A Class Mobilization Perspective"; and Leon Muszynski "Defending the Welfare State and Labour Market Policy"

6Ibid..4

7Ibid..4
impediments through such measures as reduced expenditures, deregulation, privatization, contracting out, and tightening the rules for unemployment insurance and other social welfare programs.8

Neo-liberalism, when understood as a political rationality, a mode of governing, requires the constitution of a new subject-citizen. As argued by Burchell, liberalism taken as a broad formula of rule. "...requires of the governed that they freely conduct themselves in a certain rational way."9 The 'certain rational way' is accomplished through governmental strategies which seek to constitute the subject in ways that are compatible with governmental objectives. Liberal doctrines of freedom and limits of power are thus accompanied by strategies to foster the self-regulating or self-organizing capacities of markets, citizens and civil society in desired directions.10

The integral elements of neo-liberal rationality include: a belief in the efficiency of markets, the importance of accountability, individual liberty, enterprise, responsibility and independence and the virtues of the non-interventionist state. Neo-liberal political rationality does not accept as legitimate any avoidable adult dependency. Unlike liberal welfarism, which tempered the liberal commitment to 'independence' through the employment of notions of collective social responsibility, neo-liberalism brings 'independence' back to the fore. And in contrast to liberal welfarism which tempered the devaluation of dependency by its acknowledgement of its social causes, neo-liberalism individualises and stigmatizes dependency through its reassertion of notions of individual responsibility and enterprise. 'Independence' is still fundamentally associated with waged labour and is valorized as a desirable personal quality, and dependency signifies not only lack of paid employment but "has been inflated into a behavioural syndrome and made to

8Ibid., 7


10 Nikolas Rose "Government Authority and Expertise in Advanced Liberalism" in Economy and Society Vol. 22, No. 3, August 1993
seem more contemptible."11 People who receive welfare are thus increasingly stigmatized under neo-liberalism: they are not independent, enterprising or responsible. Rather they are deviant.

Moreover, whereas the equation of waged labour with independence was gendered in the industrial period, and anxieties about economic dependency were by in large anxieties about the dependency of adult males, in the post-industrial period independence has been discursively 'de-gendered' as the industrial hegemonic acceptance of female dependency has been increasingly challenged and undermined. As stated by Fraser and Gordon. "Whereas industrial usage had cast some forms of dependency as natural and proper, postindustrial usage figures all forms as avoidable and blameworthy."12 Individual responsibility is a central tenet of liberalism. In the context of social service provision this translates into policies which have historically been based on some measure of 'deservedness': "...a person's inability to earn a living must clearly be established before public assistance is granted. The poor must always qualify as a deserving poor."13

The early 1990s witnessed a distinctly neo-liberal campaign to stigmatize all recipients of state social benefits as possible criminals through the linking and mobilization of fraud and criminality discourses in the public panic around and the governmental crackdowns on 'welfare cheaters'. A provincial government report on social welfare provision in Ontario released in 1992 in the early days of the left-leaning New Democratic Party government. drew particular attention to the problem associated with the moral determination of deservedness. It observed that the system had "...become infused with value judgements about people, about whether they are considered to be deserving or undeserving of assistance, the authors or the victims of their own


12 Ibid.323

13 Ronald Manzer Public Policies and Political Development in Canada Toronto: University of Toronto Press. 1985:50
Interestingly, at the time of this report the NDP government played down concerns about fraud and abuse of the system. While it recommended a comprehensive audit system for the welfare system, it was careful to note that this was not because "we believe there is widespread fraud in the system. We believe that it is the hallmark of any efficient system to monitor itself."\(^{15}\)

Evidencing the force of the rise in neo-liberalism as a rationality of rule, one year later the same NDP government issued yet another policy paper on social service provision, entitled *Turning Point*. The neo-Keynesian discourse of collective responsibility, entitlement, social justice and equity, so prominent in the earlier report is completely absent in *Turning Point*. In contrast, this later report gave expression to a distinctly neo-liberal discourse of competitiveness, independence, individual ‘talents’ (a specifically economic ‘value-added’ term), long-term economic strength, training, incentives and disincentives, the changing global economy and sound fiscal management. This discursive shift certainly reflects the increasing influence of neo-liberalism in political and economic policy-making. No longer do notions such as collective responsibility, human dignity and respect underpin the proposed changes. Rather the changes are justified by reference to the increasing competitiveness of global economics.\(^{16}\) Whereas the earlier policy orientation directly countered the growing assumption that social assistance is a preferred lifestyle, that welfare recipients simply don’t want to work and are abusing the system, the government in its subsequent report expresses the widely held concern that welfare recipients are “locked in a lifestyle of dependency.”\(^{17}\) The notion of a ‘lifestyle of dependence’ reflects the neo-liberal preoccupation with rational choice and cost-benefit analysis as the basis for individual decision-making about the distribution of their labour and leisure time. The implication of


\(^{15}\)Ibid., 171

\(^{16}\)Ibid., 2

\(^{17}\)Ibid., 9
this orientation for the social construction of welfare recipients is clearly that they 'freely choose' to be on welfare and therefore they are, already and always, system abusers.

III) The Emergence of the Problem of the 'Welfare Cheat' in Ontario in the Early 1990s

*Fraud:* "deceit, trickery, sharp practice, or breach of confidence, perpetrated for profit or to gain some unfair or dishonest advantage"

*Fair:* "legitimately sought...proper under the rules"\(^{18}\)

'Cheaters' offend our sense of fairness, our valorization of merit and individual accountability. In the context of humanitarian social welfare systems which are premised upon liberal notions of merit and desert, cheaters are a threat to the system. By cracking down on cheaters, the liberal moral universe is maintained and discursively reproduced.

Over the 1990's, fraudulent claims to the state's 'generosity', whether in the context of claims for refugee status or claims for social assistance have been increasingly constructed as serious social problems occasioning tough governmental action. Fraudulent claimants, 'bogus refugees' and 'welfare cheats' are the epitome of undeservedness: not only are they not deserving victims who have come to their needy situation 'through no fault of their own' but they dishonestly misrepresent themselves as victims. Moreover, in the increasingly neo-liberal context of policy-making, even the 'genuine' welfare claimant is, to a certain degree, always and already conceptualized as receiving undeserved (freely-chosen) governmental assistance.

It is telling, that in Ontario, the cutting of social assistance benefits and the crackdown on welfare fraud in 1994 was spearheaded by the New Democratic Party. The incongruity of these initiatives with the traditional politics of the NDP was not lost on political commentators. Michael Valpy, journalist for the *Globe and Mail*, observed wryly at the time, "Think of this: For the first time in Ontario's history, a provincial

government. encouraged by its premier, is talking about cutting social assistance benefits. That government is socialist. Valpy went on to observe that Ontarians saw the NDP government as ‘abandoning its promises and its socialist philosophy’ “They hear the Premier speak publicly about Ontario’s benefits being the most generous in the country. They hear him speak about the dangers of welfare dependency. On both subjects, he panders to those eager to believe that welfare recipients are lazy frauds living off the sweat of others’ labour.”

It is particularly interesting that Mike Harris, the current conservative premier of Ontario and author of the neo-liberal ‘Common Sense Revolution’in the province, articulated just this mean spirited and punitive view in his reaction to the NDP’s proposed welfare reform as then leader of the Ontario Progressive Conservative Party:

If it truly deals with welfare reform, with removing from the rolls those who are ripping us off, those who are staying home and doing nothing because they want to do nothing then that will be good enough for me...Its all those people who can get out of the house, who can do something and are choosing to stay home and do nothing.”

Since the early 1990s, the categories of ‘fraudulence’ and ‘criminality’ have become increasingly linked in social and political discourses. However, not surprisingly this is most pronounced not with corporate or white collar crime, but with respect to governance of the poor, the powerless and the marginalized. In 1995 Premier Harris stated that wealthy citizens who evade taxes, while ‘regrettable’, are reacting in accordance with ‘human nature’ against ‘wasteful government spending’ (presumably he had in mind the nature of the archetypal desirable, acquisitive, independent liberal subject). At the time of this provocative pronouncement, his government had just launched a massive ‘crackdown’ on welfare ‘cheats’ which included a toll-free provincial fraud hotline and a 20% cut of welfare rates. Opposition members were quick to point out


20 Ibid.

that in Harris’ Ontario, poor people who make fraudulent welfare claims are criminalized. They are portrayed and acted upon as crooks, whereas rich people who defraud the government are normalized, they are merely acting in accordance with their inherently acquisitive and independent (and, needless to say, essentially desirable) human nature.\textsuperscript{22}

In March 1994, Tony Silipo, the Ontario NDP Minister of Community and Social Services, announced a major crackdown on welfare fraud. The government hired 270 inspectors to review almost 690,000 welfare cases. They did this despite the acknowledged absence of any concrete evidence of the extent of the ‘problem’: ‘Mr. Silipo said he does not know just how serious a problem welfare fraud is...[He] said the government is responding in part to a widely held belief that welfare fraud is increasing, even though there is no evidence to back up that belief.’\textsuperscript{23} Indeed, what evidence does exist places the level of welfare fraud at less than 3% of all claims.\textsuperscript{24}

This ‘crackdown’ and the admittedly unsubstantiated ‘problem’ of welfare fraud that it targeted received massive media coverage. A brief survey of the headlines in \textit{The Toronto Sun}, a conservative daily tabloid newspaper illustrates the tone and mood of the public panic. Headlines screamed: “Living High off the Hog”\textsuperscript{25}, “Cheaters Beware”\textsuperscript{26}, “Welfare ‘Gravy Train’ Derailed”\textsuperscript{27}, “Welfare Fraud Deluge”\textsuperscript{28}. The extent to which the ‘problem’ of welfare fraud was constructed and manipulated is further revealed in confidential documents obtained under the Freedom of Information Act which were

\begin{itemize}
\item \textit{Globe and Mail} “Tories Accused of Double Standard on Fraud: Welfare Cheats Face Courts as ‘Crooks’ While Tax Evaders get Sympathy for being ‘human’ Liberals Charge” Nov.16. 1995
\item \textit{Globe and Mail} “Ontario Takes Tough Stand on Welfare Cheats” March 29, 1994
\item \textit{Toronto Star} “Welfare Bashing” April 3, 1994
\item \textit{Toronto Sun}. March 18, 1994
\item \textit{Toronto Sun}. March 28, 1994
\item \textit{Toronto Sun}. April 14, 1994
\item \textit{Saturday Sun}. February 19, 1994
\end{itemize}
presented in the legislature in April 1994, but which did little to diffuse the panic. These documents attributed the 'skyrocketing' costs of provincial social assistance to government error in the form of massive overpayments. As put by Liberal leader Lyn McLeod. "Instead of hiring 270 inspectors to search for welfare cheats the government should get its own house in order and stop sending excess money out to people...The problem is really government mismanagement."29

As part of the 1994 crackdown, the NDP also investigated the idea of issuing welfare ID cards to social assistance recipients as a means of combatting fraud. As well it even looked into fingerprinting as an anti-fraud measure.30 There is perhaps no greater illustration of the criminalization of welfare recipients underway at the time than that provided by the idea of taking their fingerprints. entailing as it does "...the physical laying on of hands that fingerprinting requires, the taint of criminality that has always surrounded it, the sense of being psychologically violated..." Larry Hannant warns. "Government officials who want to impose it had better face up to history: Barring a national emergency - and the supposed epidemic of welfare fraud hardly qualifies - Canadians won't leave their mark with the government or police."31

IV) The Merging of Fears about Immigration and Welfare

1994 was also a big year for controversy and debate about Canadian immigration and refugee law and policy. The federal government's announcement in February that it was going to maintain immigration levels at 250,000 sparked much criticism. Summing up the basic objection, the Reform Party of Canada said "...immigration is still twice what it should be ... demanded an explanation for continued high levels while unemployment

29Toronto Star "247 Million Was Overpaid on Welfare, Liberals Charge" April 14, 1994

30Toronto Sun "Double Duty for Photo ID Cards", April 21, 1994:16; Globe and Mail
 "No Fingerprints. Please. We're Canadian" March 7, 1994

31Globe and Mail Larry Hannant, "No Fingerprints, Please. We're Canadian" (op-ed.) March 7. 1994
continues to soar and Canada's social welfare system is under strain." In its anti-immigration platform, the Reform Party routinely alluded to "rising crime rates, bulging welfare rolls, unemployment and underfinanced social services."

Support for the arguments of Reformers and other anti-immigration advocates is provided by freelance Toronto journalist, Daniel Stoffman and author of a report on immigration for the independent conservative think tank, the C.D. Howe Institute.

Stoffman refutes the argument that immigrants are an economic asset to the Canadian economy, arguing that their effect on the incomes of the 'host' country is 'neutral'. For this reason, he suggests that economic justifications for existing immigration policy are not well-founded. Instead, Stoffman draws attention to such non-economic deleterious effects of immigration as increasing racial tensions and the overburdening of social services. For these social reasons, Stoffman concludes that immigration should be reduced and that the priority should be placed on taking more skilled and highly trained and educated independent applicants and fewer humanitarian ones.

---

32 Toronto Star. "Liberals Open Canada's Door to 250,000 immigrants in 1994" February 3, 1994

33 Toronto Star. "Nose to Nose on Immigration" February 5, 1994

34 Globe and Mail. "Putting a Price on Immigration" February 11, 1994

35 The February 1994 announcement that the government was not going to reduce immigration levels was met with dismay from anti-immigration groups who had hoped, in line with Stoffman, that they would be. However, the changes that the government did make to the intake levels of the different classes of immigrants reveals a certain agreement with the view articulated by Stoffman and others. These changes do indeed reflect certain changes in priority with respect to the admission of different classes of immigrants. Within the overall number of 250,000, there were to be more independent immigrants chosen for their skills and potential economic contribution to the economy, more family class immigrants and significantly fewer refugee admissions. In this last respect, off-shore refugee selection was preferred. This reflects another interesting analytical shift in contemporary inclusionary notions of governmental humanitarianism which have been increasingly associated with 'families' and 'children,' and decreasingly with refugees. In 1994, approximately 45% of those immigrants were to be accepted under the family class category, 44% were to be selected on the basis of the point system.
opens the door for positions on immigration which emphasize the alleged deleterious effects of immigration, and in particular of 'uncontrolled' on-shore refugee admissions, including as social decay, racial tensions, cultural difference, safety and security issues. Such arguments easily provide discursive cover for underlying racist and xenophobic sentiments.

Over the late 1980s and early 1990s, the 'problems' posed by immigration were the subject of many polls and surveys which sought to uncover Canadians' opinions about immigration. The findings consistently revealed a Canadian public that was increasingly intolerant of immigrants and critical of immigration policy. Many of these studies found Canadian opinions on these matters to be inflected with xenophobic and racist fears and beliefs.\(^6\)

So it was that by the early 1990s, the regulatory figures of the 'bogus refugee' and the 'welfare cheat' were well-established in Canadian public and political discourse and efforts were underway to crack down on both. Over this period, fraudulent claims for social benefits were increasingly criminalized. However, as already suggested, this criminalization was not felt equally by all Canadians. Corporate fraud, white collar fraud, tax fraud were not the subject of political and public panics over this period. Rather, it was the least powerful members of society who were targeted: new immigrants and refugees and welfare recipients.

---

for skills and business credentials and 11% were to be accepted as refugees fleeing persecution. (The Toronto Star, "Nose to Nose on Immigration" February 5, 1994) The decline in refugee admissions is telling, particularly at a time when the world refugee population was increasing dramatically indicating a shift in priority within the humanitarian class of immigration admissions from refugee admissions to family class admissions.

V) Extending and Acting Upon the Threat: Cracking Down on the ‘Fraudulent Criminal Claimant’

The increased preoccupation with and criminalization of (certain kinds of) fraud under neo-liberalism facilitated the relatively easy link forged between fraud and criminality mobilized against new immigrants and refugees. Neo-liberalism, premised as it is upon notions of individual responsibility, autonomy and rational choice, has a distinctly punitive edge. Thus while criminalization is not necessarily consistent with neo-liberal strategies of governing, implying as it does increased costs of criminal justice administration and state administered carceral institutions, the language of deterrence central to criminal justice processes is nevertheless consistent with the neo-liberal preoccupation with utilitarian notions of rational choice and cost-benefit decision-making. The logic of deterrence, in the context of immigration just as in the context of criminal justice or social welfare, justifies an enforcement oriented, get tough, law and order response to whatever ‘threat’ is being posed. By increasing the ‘costs’ of coming to Canada through punitive crackdowns, ever tougher legislation, more rigorous enforcement, interdiction initiatives and increased sanctions, the rational decision-maker will be dissuaded from choosing to come to Canada to abuse the system and endanger the public.

One of the troubling aspects of the increasingly deterrence-oriented Canadian immigration law and policy is that it tends to construct all new immigrants and refugees as system abusers and criminal threats, who in turn are acted upon as such. Thus it fosters a ‘culture of disbelief’. While ‘genuine’ refugees, by legal definition, are still widely considered and are acted upon socially, legally and politically as deserving victims, the vast majority of refugee claimants are represented as criminals or system abusers or both. While it is generally conceded that genuine refugees are deserving of protection and that genuinely needy welfare recipients are deserving of assistance, the discursive emphasis has shifted to the undeserving and undesirable side of the liberal equation.

However, Canada’s relatively high rate of accepting refugee claimants as genuine
refugees reflects a judgment that most claimants are not frauds or criminals despite their current construction. Under the current governmental regime, the legally and administratively confirmed status of the refugee claimant as a genuine and therefore deserving refugees is fragile and arguably always suspect. In one moment, and in a wide range of social, political and legal contexts, the claimant is constituted as, in all probability, an undeserving, opportunistic criminal, fraud or both, who ‘freely chooses’ to come to Canada for illegitimate reasons, and who poses a threat to national security, protection of the public and/or system integrity. In the next moment, legally and rhetorically the successful claimant is confirmed to be a deserving victim who had no choice but to flee to Canada and who therefore has a right, under national and international law, to be here.

VI) The Provincial and Municipal Dimensions of the Panic Surrounding the ‘Fraudulent Criminal’ Immigrant or Refugee

Federal developments attest to the heightened preoccupation with and discursive links between immigrants and/or refugees, criminality and system abuse at this time. Provinces and municipalities were no less animated by the same issues. The level of fear-mongering and hyperbole which similarly characterised the more local panic in Ontario is nicely captured in the following dire picture of the state of Canadian immigration painted by a Toronto Sun columnist in 1994:

Worst of all the word has gone out to foreign violent criminals that Canada is easy to enter and has a loosely run system that can be ‘worked’ to frustrate deportation even when caught. Many of these criminals—murderers, drug traffickers, terrorists, pimps, etc., - apply for “refugee” status and stay on for years. And when they run out of time they simply

Indeed, arguably provinces and municipalities were motivated by a further issue not shared at the federal level, that is the rising costs borne by provinces and municipalities for social service provision due to cuts in transfer payments. In addition, Metropolitan Toronto has always been extremely vocal on immigration issues due to the huge numbers of new immigrants and refugees who choose to live there and the particular issues that that raises as well as due to the potent dictates of electoral politics.
When you see the amount of chronic, large scale unemployment in Canada - with 3 million on the welfare rolls - you have to wonder about the insanity of these new immigration increases. 

The above quotation provides a compact compendium of the dominant discourses mobilized in the governance of immigration at this time: criminality, risk/danger, fraud, system abuse, and their association with economic insecurities and welfare. It is through the operation of these discourses that the figure of the deserving refugee is transformed and undermined. Panics tend to focus on simplistic identities and here, the deserving victim, the genuine refugee, is consistently eclipsed by the fraudulent criminal.

In April 1993, restricted police and immigration documents were leaked to a municipal politician in Toronto, John Papadakis. Soon after, Papadakis released them at a public forum on crime which he had organized after seeing a report issued by the East York Council on the impact of refugee claimants on local municipal governments. In attendance at the forum were other municipal politicians, representatives of the Immigration department and of the police. Its stated purpose was to “focus attention on bogus refugee claimants and illegal immigrants”\(^\text{40}\). The fact that they were confidential leaked documents only bolstered Papadakis’ resolve. As he explained, the fact that individuals would leak this information is proof that ‘law-abiding’ citizens are “fed up with government inaction, with having their country abused, with being afraid to walk around at night, with being robbed in their stores, and with being accused of being racist every time they open their mouths to say there’s something wrong”.\(^\text{41}\) Another reading is possible of the bureaucratic leak which Papadakis acted upon. As observed by Lorne

\(^{38}\)Toronto Sun. “Immigrant Boost is Insanity” Feb.4, 1994

\(^{39}\)Globe and Mail. January 21, 1993

\(^{40}\)Globe and Mail “Probe into Leak on Refugees Sought”, April 29, 1993

\(^{41}\)Ibid.
Waldman. Member of the Refugee Lawyers' Association.

A broader question has to be asked: Who's leaking this and why? Some people inside the Immigration Department have an agenda... This agenda a lot of times, is an agenda to try to create an anti-immigrant, anti-refugee backlash. There are refugees who collect welfare... There are visitors who come to Canada who commit crimes. But as a proportion of the total number of people who come to this country it's very insignificant.\(^{42}\)

Papadakis, and his allies in Immigration, used this information to further forge the link between illegal immigrants and refugee claimants and the 'crime problem'. This link, as already described, had been evidenced by the creation of the new 'Public Security' portfolio in the same year thus reproducing the discursive message that immigrants and refugees are a threat which needs to be contained and controlled: "...a menace that society has to be protected from."\(^{43}\)

Papadakis emphasised that national security and public safety supersede any privacy laws: 'every Canadian has a right to safety'. The easy slippage between criminality and fraud/system abuse, as well as the general view that any government assistance is, in a sense, 'criminal', is evidenced in his remark that "Crime is Crime. When my folks came to Canada in the 1950s they had no help. no handouts."\(^{44}\)

The leaking of these documents by civil servants, their release at a public forum and the media reports on this event all contributed to the continued production and reproduction of the panic around immigrants and refugees, fraud/system abuse and criminality. The information contained in them was presented by Papadakis as "proof" that "Canada has a serious problem with foreigners coming into the country and

\(^{42}\) *Globe and Mail* “Use of Leaked Data Called Irresponsible”. April 30, 1993

\(^{43}\) *Globe and Mail*. “Shifting New Ministry seen as Fanning Flames of Racism July 13, 1993

\(^{44}\) Ibid.
committing violent crimes'.

He concluded that any and all refugee claimants or immigrants who are convicted of a criminal offence should be automatically deported, with no right of appeal.

The linkage between immigrants and refugees, fraud/system abuse and criminality discursively constituted all new immigrants and refugees as potential threats to national security, public safety and system integrity. However, it was also mobilized in local, 'ethnically-specific' directions. In Toronto in the 1990s, Somalis were one group which bore the coercive brunt of this development. Their experience illuminates the nature of these discursive processes and their coercive impact on people.

VII) From ‘Deserving Victims’ to ‘Fraudulent Criminals’: The Case of Somali Refugees in Toronto in the 1990s

The historically specific conceptions of (un)desirability and (un)deservedness which underpin and guide the development of exclusionary immigration law and policy and the dominant rationales that explain and justify their enforcement are constituted, mobilized and reproduced by and through the discursive ‘production’ of particular social problems (dangers, threats, risks) and by and through efforts to regulate and control the ‘problem’. In the production of these threats, discourses are mobilized and employed: criminality, risk/danger, fraud, abuse, system integrity, Canadian generosity and sovereignty are but a few which have been highlighted here.

In what follows, the concrete and coercive impact of these discursive developments will be explored in the particular context of Somali refugees residing in Toronto in the 1990s. The processes by and through which ‘genuinely deserving’ Somali refugees were incrementally and steadily reconstituted and acted upon as undeserving and undesirable will be examined. It is argued that this redefinition was effected primarily through the mobilization of fraud and criminality discourses against particular groups of

---

45*Globe and Mail.* “Use of Leaked Data Called Irresponsible” April 30, 1993

46*Globe and Mail.* “Probe into Leak on Refugees Sought” April 29, 1993
new immigrants and refugees who were (re)presented as 'threats' to Canadians and
Canada. What follows makes clear that these discursive developments entail concrete and
coercive consequences for those subject to their operation. In the case of the Somalis,
these consequences reached quite literally into their very homes.47

This discussion is not intended as a comprehensive micro-study of the production
and consequences of the anti-Somali panic and backlash in Toronto in the mid-1990s.
although that would most certainly be a valuable and interesting project to pursue. Rather,
what follows provides an overview of that situation as illustrative of the convergence.
employment and racist and coercive dimensions of fraud and criminality discourses in a
local and human context. Key players in the multi-contextual production of the panic and
the coercive crackdown that ensued included: ‘Dixon’ condominium owners, security
personnel and residents/owners; the city of Etobicoke Social Services ‘Dixon Task force’;
the print and broadcast media (in particular the Toronto Sun. and an article entitled
“Dispatch From Dixon” by Toronto freelance journalist Daniel Stoffman which appeared
in the Toronto Life Magazine in August of 1995), municipal and provincial politicians.
Immigration ‘intelligence’ bureaucrats, the RCMP. and the municipal police. Each
‘player’ could easily be subjected to much more detailed and thorough description and
analysis.

We begin with a quotation from an Africa Watch report on the distressing
situation in Somalia by the beginning of the 1990s:

47

In the final section of this chapter, this linkage will be examined in the context of the
Somali community in Toronto. It should however be emphasised that other groups were
and continue to be similarly governed by these discursive developments. for example the
Sri Lankan community felt the effects of a widely covered emergent concern with
criminal ethnically based gangs as well as with fundraising activities in support of
‘terrorist’ activities in their homeland; the Chinese community was effected by a
sustained and widely covered panic about ‘Triads’ (the ‘threats’ posed by organized crime
groups and measures taken to respond to this threat will be examined in Chapter 6); Sikhs
were the subjects of considerable enforcement-oriented fall-out after the Air India crash
in 1985; and people from the Middle East felt the effects of the sustained enforcement
preoccupation with (fundamentalist) terrorists.
It is difficult to overstate the Somalia government's brutality towards its own people, or to measure the impact of its murderous policies. Two decades of the Presidency of President Siad Barre have resulted in human rights violations on an unprecedented scale which have devastated the country. Even before the current wars the human rights of Somalis were violated systematically, violently and with absolute impunity.\(^{48}\)

Over the 1980s and 1990s, hundreds of thousands of Somalis were forced to flee state sanctioned persecution in their country. Most ended up in neighbouring countries in refugee camps. However, many who had the resources to travel further afar, were granted refugee status in Canada and the United States. The Somali community in Toronto began to form in the mid 1970s soon after refugees had begun to flee the Ethiopian-Somalian war. Thousands more followed over the 1980s and 1990s as conditions within Somalia steadily deteriorated. According to the Somali Immigrant Aid Association, over the 1990s the size of the Somali community in Canada grew from 65,000 to over 90,000: “Next to the Sri Lankan Community, ours has one of the largest influxes of desperate refugees seeking refuge... in Canada.”\(^{49}\)

After arriving in Canada through the major airports in Toronto, Montreal, Vancouver and through the Canada/US border, many Somalis moved to Toronto, drawn by the large Somali community. Thousands of Somalis live in six high rises on Dixon Road in Etobicoke, a suburb of Toronto.\(^{50}\) In the general panic that had been created about the character and fraudulent and criminal activities (identities) of Somali refugees, conditions faced by the Somalis living at ‘Dixon’ steadily worsened as the more general panic most certainly fuelled and propelled a racist and coercive backlash which was played out where they lived.

\(^{48}\)Africa Watch *Somalia: A Government at War with its own People*, New York: 1990:1

\(^{49}\)Somali Immigrant Aid Association Website, Background Information, October 1999

\(^{50}\)Hereafter referred to as ‘Dixon’
i) Somali Refugees: ‘Cheaters’, ‘Criminals’ and ‘Unscrupulous Masters of Confusion’

This ‘ethnically-specific’ tale of the mobilization of the merged threat posed by fraudulent criminal claimants begins with yet another leak of ‘secret federal reports’. this time to provincial liberal leader Lyn McLeod in October, 1993. These reports, one of which bore the racist and inflammatory title “Desert Gypsy”, were written by a bureaucrat in the Immigration Department’s Intelligence Unit. They alleged that Somalis were engaged in widespread and organized welfare abuse and that the Canadian government was being defrauded of millions of dollars as a result of this abuse.

McLeod’s interest in the reports was not in the evidence they supplied of the prejudices within the Immigration Department. Rather she used them as part of her campaign against welfare fraud. In the reports, Somali refugees were described as “masters of confusion” who “were importing refugees to systematically pillage our vulnerable and exposed social welfare systems.” The report also contained allegations that the money was being used to buy weapons for use by Somali ‘war lords’ in the continuing civil war in Somalia. The report, according to another source, went on to describe Somalis as “the greatest threats to Canada’s internal security.” Somalis were described as ‘opportunists’ whose “...use of confusion and misrepresentation is unparalleled except by the Gypsies of Eastern and Western Europe.” As described in a *Toronto Star* article, the author of the report claimed that “....Our western and primarily Christian-based way of life has little meaning or relevance to these people...one Somali interpreter has been quoted as saying you can

51 This title gives blatant expression to the increasing prejudice against traditionally nomadic people over this period. This prejudicial theme, as will be seen, resurfaces in the context of the Somalis, but which clearly peaked in reaction to the arrival of ‘Gypsy’ refugees in Canada in the mid-1990s.

52 *Toronto Star*, “Minister Unveils Steps to Curb Welfare Abuse” March 10, 1994

53 *Globe and Mail*, “Refugees Accused of Fraud” October 28, 1993

54 *This Magazine*, “They Believed the Hype” Dec-Jan. 1995:30
believe only 50 per cent of what a Somali tells you. My experience tells me that 50 per cent is extremely generous.”

The substance of these allegations were also reported in a Vancouver Sun article in October 1993. This story relied in large part on comments made by an anonymous former immigration investigator whose allegations were reportedly confirmed by similarly anonymous ‘law-abiding’ Somalis. It reported that tens of millions of dollars in welfare money was being fraudulently collected by Somali refugees across Canada and that the practice of ‘tithing’ resulted in a huge percentage of this money being sent back to Somalia to buy weapons for ‘warlords’. The article quotes the words of this former investigator who, when asked why Somali refugees come to Canada when many of them arrive in the US first, responded that ‘You can’t buy arms in Mogadishu with food stamps’.

This report was leaked and publicized at the time of Silipo’s crackdown on welfare fraud in Ontario. While distancing himself from the specifically racist dimensions of the allegations, Silipo did little to calm the fears about refugees and welfare fraud. He responded that the government was currently investigating allegations that ‘some refugees are receiving multiple welfare cheques by using false identification.” While he observed that some of the statements in the report could incite racism, he nonetheless did not challenge the validity of its content: “I don’t know if the particular report is authentic. I can’t tell, but certainly the issue that’s raised is an authentic issue.”

Just as the more general ‘welfare cheat’ panic was not grounded in any statistical evidence of the actual extent of the problem, this refugee-specific panic was also not substantiated by any numbers. Numbers are indeed irrelevant in these panics. As admitted by Silipo, the government of the day had no estimate on the size of this ‘problem’: “I

55 Toronto Star. October 29, 1993
57 Globe and Mail; “Refugees Accused of Fraud” October 28, 1993
58 Ibid.
don't think that one can say there is fraud in the millions. I don't think we know that quite honestly...I think that what we have to do is to make sure we get at whatever level of fraud there is in the system.” In fact, according to Silipo, the report in question provided “very few” cases of fraud, “something like twenty.” Moreover, according to a subsequent news story, “...the reports make specific reference to only a handful of cases and provide no proof - other than assurances from Somali community ‘sources’ - that money being bilked from welfare is being sent to warlords in Mogadishu. Thus while the focus of the allegations was on Somali welfare fraud, this report raised the added dimension of criminality evoked by the reference to welfare payments for weapon purchases in Somalia. The confirmed legal status of Somali refugees as deserving victims was thus unsettled. Somali refugees were reconstituted as undeserving and undesirable threats. Indeed, their engagement in ‘systematic’ welfare fraud, already socially and politically criminalized, was said to be directly in aid of further ‘criminal’ activities abroad.

The leaked document and the allegations it contained received very wide coverage in the press. The very identity of Somalis was being recast. Instead of deserving victims of persecution. Somalis were re-presented as unscrupulous and cunning ‘masters of confusion’: they were moreover seen not as not novices or newcomers to this kind of activity - but rather they were pros, ‘masters’. While the report itself was subsequently discredited in the media and even the Minister of Immigration, Sergio Marchi, explicitly distanced himself from it, the dye had already been set.

The damage, which was severe, had been done. Shortly after the report became public. 500 Somalis met in Ottawa to discuss their situation and many reported

59 Ibid.

60 Toronto Star.”Minister Unveils Steps to Curb Welfare Abuse” March 10. 1994

61 Bill C-86, it will be recalled, had ‘criminalized’ (through its inadmissibility and removal provisions) any activities of individuals which could be linked with terrorism, organized crime or support of groups, governments or organizations that use violence and/or engage in human rights violations.
experiencing the coercive edge of this development; according to one report, "many told stories of harassment by neighbours and strangers, of being denied jobs and of their children being taunted at school."^62

Over 1993 and 1994 this 'fraudulent criminal' identity of Somalis was the particular object of media attention. In 1994, the *Toronto Sun* tracked a number of welfare fraud allegations, taking care to specify both the alleged offender's ethnicity (Somali) *and* immigration status (refugees or refugee claimants). Headlines were sensational. Even mere 'suspicions' of welfare fraud by Somalis were reported at this time: "A Somali refugee claimant is being sought on suspicion that he defrauded an immigration welfare program of $4000 while his claim was being heard."^65 Another article entitled "Somali Woman Shopping for a Country"^64 included a 6" by 8" photograph of the woman in question and her two children, identifying her in upper case letters as "COLLECTOR". While in the text the woman is later identified as having claimed refugee status, she is introduced as "Somali expatriate". By avoiding the designation "Somali refugee", her national origin is emphasised while her status as refugee claimant is not acknowledged; even the merest possibility of evoking her 'victim' status (and hence identity) is completely avoided. On March 16th, 1994 the *Toronto Sun* ran another story on the same woman entitled "On the Move on the Dole"^65. Once again, the woman is identified without mention of her refugee status. this time she is a 'Somalia-born globetrotter'.

The latter designation speaks to the 'problem' of 'asylum-shopping' which had also been the subject of increasing attention during this period. Indeed a major official justification for the restrictive and discriminatory safe third country provision of Bill C-

^62 *This Magazine*. "They Believed the Hype" Dec-Jan 1995:30

^63 *Toronto Sun* "4Gs Welfare Payout Probed" February 21, 1994

^64 *Toronto Sun*. March 11, 1994

^65 *Toronto Sun* "On the Move on the Dole" March 16th, 1994
86 was the need to stop ‘asylum shopping’. Asylum shoppers is the derisive term used for those suspected of “freely choosing” to come to Canada rather than being compelled by virtue of state sanctioned persecution. People who ‘asylum shop’ have been, or could have been granted asylum in a so-called ‘safe’ third country in which they had been present before arriving in Canada. Oddly, justifications like the need for western ‘developed’ countries to “share the burden” of the world’s ills were mobilized in defence of this restrictive measure at a time when most other western European countries were doing the same thing: closing their doors and thereby shifting, not sharing the burden.

While asylum shoppers may once have been recognized as ‘genuinely deserving victims’, this status is negated as soon as they ‘freely choose’ to leave the safe place which had ‘generously’ and ‘compassionately’ allowed them to stay, and opportunistically and undeservedly claim refugee status in Canada. Asylum shopping, like the ‘bogus refugee’, offends liberal sensibilities. And while not to the same degree perhaps as the ‘bogus refugee’, during this same period ‘asylum shopping’ was similarly constructed and acted upon as an emergent and serious ‘problem’ requiring restrictive and enforcement oriented attention.

Despite the absence of any reliable statistical evidence of the extent of the ‘problem’ of welfare fraud (nor indeed of any reliable ethnically-specific data on this issue) and despite official pronouncements on the exaggerated reaction to this so-called problem, Immigration Minister Sergio Marchi announced in March, 1994 that ‘information sharing’ measures were being implemented with municipal governments in order to assist in the crackdown already in progress. Once again, Marchi announced these initiatives while being at the same time cautious about exaggerated reports of the problem: “...the problem of welfare abuse by refugee claimants - while worthy of attention and action - should not be exaggerated.” He also took the opportunity to again distance himself, as did in November of 1993, from reports about Somali welfare abuse: “I just don’t buy it when there’s references in the report that go from individual guilt to a
collective community blanket guilt.”  

The extent to which the identity of Somalis became indelibly tainted by unsubstantiated welfare abuse and criminality related allegations was evidenced by the common-place association made between welfare fraud and Somalis. In an April 1994 Toronto Star piece critical of the ‘scapegoating’ of welfare recipients that was occurring in Ontario over this period, the association between welfare fraud and Somalis which is increasingly taken for granted is highlighted: “You hear everywhere a little litany that goes something like this:...A friend of mine told me about this Somali guy who came right from the airport to the welfare office.”  

As Don Richmond, the Commissioner of Metropolitan Toronto’s Social Services noted. “[W]elfare clients are becoming the other, the stranger, the Jews.”

In 1994, similarly unsubstantiated allegations about the fraudulence and criminality of Somalis surfaced in the local community in Etobicoke. According to Ali Mohamud, at the time a member of the Somali Community in Etobicoke and presently (2000) Executive Director of Dejinta Beesha, a Somali settlement services community organization, members of the Somali community were accused of misappropriating government funds that had been allocated to Dejinta Beesha. As with the Somali welfare fraud allegations, these local allegations also linked fraud and criminality. Somalis were accused of redirecting the fraudulently obtained funds to Somalia to support the ongoing criminal activities of warlords.

This allegation triggered an RCMP investigation. In 1994, members of the RCMP came to Dejinta Beesha to investigate the charges. According to Mohamud, the RCMP was satisfied with the legality of the organization’s operations and produced a report

---

66Toronto Star. “Minister Unveils Steps to Curb Abuse” March 10, 1994
67The Toronto Star. “Powerless welfare clients are Scapegoats of the 90s” April 8, 1994
68Ibid.
which indicated that there was "no substance to the allegations." This was however, not the end of the matter. As will be shown, less than two years later the very same allegations sparked a second RCMP investigation. Its findings were the same as those in 1994.

The panic which developed around the 'problem' of cheaters and criminals in the early 1990s in the context of the administration of Canadian social welfare and immigration policy converged upon the Somalis in a particularly powerful and explicit way. The panic that grew around the Somalis was facilitated by a national social and political climate of rising fear of and hostility toward outsiders, economic insecurity and a certain lack of confidence in government. It was also fuelled by, and in turn helped to fuel, racist and xenophobic sentiments. For Somalis living in Toronto, it most certainly contributed to an increasingly difficult situation in which they struggled in various ways to be heard in the face of increasingly oppressive preconceptions and repressive practices.

ii) Conflict and Coercion at Dixon

Thousands of Somali refugees live in six high rises located on Dixon Road in Etobicoke (Dixon). The population of Somalis is so great here that these buildings are sometimes referred to as Little Mogadishu. Many family members still in war-torn Somalia will know about Dixon from their correspondence with those who managed to leave. Many of those left behind continue to wait to be reunited with their families here. Foiled by the twin administrative requirements that permanent residency not be granted without 'satisfactory' proof of identity and that refugees cannot sponsor family members until they are permanent residents.

In the early 1990s the numbers of Somalis living at Dixon increased, commensurate with the growing exodus of refugees from Somalia. In July 1993, hundreds of Somali residents demonstrated on Dixon Road in protest of what they described as racist and discriminatory treatment by building management and security. According to

69 Interview with Ali Mohamud, Executive Director, Dejinta Beesha (Somali Multi-Service Centre) January 13, 2000
press reports and subsequent interviews with Somali and non-Somali residents. This demonstration was triggered by an altercation over a parking infraction between a security guard and a Somali visitor to the building. As the altercation developed, a crowd grew and the security guard called for reinforcements. A security dog was brought onto the scene by the reinforcements. It was let loose by the guards and soon after bit a Somali woman who had to be taken to hospital for medical attention.

Interviews carried out with resident Somalis and non-Somalis in 1994 by Kelly Grover reveal a history of tensions and conflict among some Somalis and non-Somalis and between some Somalis and security personnel dating back to 1990. The interviews clearly point to overcrowding as a central variable in the hostilities. However once again, it is not just about numbers. The interviews reveal that for many non-Somali residents, the problem of overcrowding was inextricably connected to and heightened by the perceived 'problem' posed by Somalis in particular. Building security personnel were frequently called upon by disgruntled non-Somali residents to take action against Somali residents or visitors whom they regarded as ill-behaved and building security frequently dealt with Somali residents in particularly hostile and coercive ways.

Much of the tension that had been brewing before the demonstration of July 1993 can be explained by a lack of communication and understanding between residents. However, the nature and degree of the hostilities was certainly heightened by distinctly racist sentiments, no doubt fueled to varying degrees by the more general denigration of the character of Somalis that was taking place around them. It is suggested here that these hostilities, and the conflict and coercion which they entailed, were facilitated by the

70 I am grateful to Kelly Grover for sharing her data with me. The interviews she carried out with the residents of Dixon were used in her Masters' thesis: "The Social Organization of a High-Rise Neighbourhood: The Influence of Race, Culture, Economic Class and Tenure on the Community Sentiments of Kingsview Park." Grover's thesis was for the School of Urban and Regional Planning, Queen's University. June 1995.

71 Ibid. All subsequent quotations of and observations about the residents of Dixon and are drawn from the transcripts of the interviews carried out by Kelly Grover.
already familiar and well-entrenched social and political construction of the undeserving and undesirable fraudulent criminal refugee. In a sense, this broader discursive development provided a certain moral legitimacy and justification for the hostility of non-Somali residents towards the Somali residents and for the coercive and discriminatory treatment of Somali residents by security personnel. It is therefore reasonable to suggest that the reactions of Somali residents to such hostilities and coercion were also likely heightened by their awareness of the distorted and offensive picture that was being painted of them in political and public discourse and their desire to challenge and resist it.

Non-Somali residents at Dixon describe Somalis refugees as ‘animals’ who have no respect for rules or for non-Somali residents. They talk too loud, gather in the courtyard in large groups, they don’t control their children, they walk on the grass, they don’t hold the door open for non-Somali residents, they monopolize the elevators. Totally unsubstantiated rumours flourish about Somalis defecating and urinating in the elevators. They routinely flaunt parking regulations and other rules of the apartment complex. They leave garbage in the hallways. They don’t talk to anyone but themselves. They are rude, aggressive and impolite. They are unclean and uncivilized. Because they are ‘transient’ renters and not owners, they have no respect for the upkeep of the building and residential units. They sit around all day and collect welfare. They own cars.

Evidently, the problem is not merely overcrowding. While there is certainly no question that many of the complaints of non-Somali residents were not racially specific and referred to problems that clearly stem from overcrowding: rising utility costs, increased noise, increased usage of the elevator and few parking spots. more often than not, the identification of these more practical issues was infused with denigrating moral and racial stereotyping. For example, one non-Somali resident commented that the major problem at Dixon was the problem of overcrowding, and that management needed to limit the numbers of residents. The moral underpinnings of her position were quickly clear when she stated that only single families should live in the units. “...no opposite sex kids in the same bedroom. Two of each sex in a bedroom that’s a normal family. A normal family is not a clan.”
Another respondent who was also dismayed by the problem of overcrowding referred to increased elevator usage as a negative consequence. However the manner in which she articulates this issue is heavily loaded. The elevator is more crowded, she observes, because "...if they have to go up two floors they will take the elevator." The implicit character judgment here is that Somalis are lazy. The same respondent observed that overcrowding has entailed more people traffic in the courtyards. As she put it, "there are more kids under ten, running all over". The problem, she suggested, is with the poor parenting skills of Somalis who do not supervise their children adequately. Thus the "problem" which elicits the most moral outrage and hostility is not overcrowding per se, but rather with the imputed (im)moral character of the Somalis.

Some of the non-Somali residents were remarkably candid in their remarks about Somalis. One resident commented that "these people were as low as animals could go". The same respondent observed that while there had always been a racial and ethnic mix at Dixon, prior to the Somali "invasion" everyone had gotten along well. Another respondent observed, "...this is a country for humans and when you bring animals in from other parts of the world, it takes animals time to adjust...these people are not working, are on fixed incomes, and are sitting at home."

Doubts were raised about the credibility of the Somalis refugee status. There is little question that in the eyes of many of the non-Somali residents, the Somali residents of Dixon did not fit the dominant construction of the genuine and deserving refugee. They did not appear to be grateful to Canada and Canadians for their generous protection. They were "aggressive" and "difficult", rather than compliant. They were wealthier than refugees should be and, to make matters worse, they "milk the system" by collecting welfare. One respondent observed that "...all of a sudden the fad is when refugees come they can afford cars - when I was an immigrant we had to work for many years to afford a car. As a matter of fact since my husband died I don't have a car. Refugees are getting free rent and free homes, we're not...what do we do with our lives, we can't start over." And most frequently it was reported that they were flagrant 'rule-breakers'. They were, in essence, undeserving and morally deficient.
A few of the non-Somali residents who were interviewed expressed dismay at the way in which Somalis were being represented and treated by non-Somali residents and by security. One respondent noted that in her opinion, "...management, security and racist owners have caused a lot of the problems in this community." Another observed that he didn’t think that security was "...very good with Somalis in the courtyard...I have watched and they pick on them. Somalis just want to sit out there." Another respondent stated bluntly that she had "...no respect for this administration or the in-house security." In her view the security personnel were racist.

Non-Somali residents offered a variety of similarly 'get tough' recommendations to ease the 'problem' at Dixon. These included: tighter immigration screening procedures: increased removals ("people should be kicked out of Canada if they misbehave"): limits on the numbers of people allowed to live in the units: stronger enforcement of the rules: a prohibition on renting to welfare recipients ("get rid of welfare recipients"): facilitate management’s access to the units (to "get proof of what's going on"). distribute the Somali residents more thinly ("...even out the population. spread the one dominant culture around so that you don’t have a concentration of one group.")

While a few non-Somali respondents focussed on the lack of understanding and communication between Somalis and non-Somalis. Somali respondents invariably emphasized the importance of encouraging communication and cross-cultural understanding between the two groups. Many advocated holding workshops and seminars to facilitate communication and understanding. Many felt misunderstood. their identity and experiences as refugees unknown, forgotten or ignored. The vastly different conceptions held by refugees. and Somalis in particular. of what constitutes a 'problem' is emphasized here:

Some other residents might say there is noise such as the aeroplanes which are a problem but when you have stresses like you have. like the problems in Somalia - you don’t notice these things as much. People from Somalia had other pressures when they were living in Somalia - people dying. trying to stay alive.

People don’t understand how hard some of these people have it. A lot of
Somali women...now have no husband, they don’t speak much English. They have five or six children. They are in a dream, a trauma. People don’t understand. There is a lack of communication. Neighbours don’t understand. They think that all of us want government help and want it easy. Many of us want to work and can’t find daycare. These women are not having it easy and people don’t understand because of language barriers and lack of communication. Not all of us are on welfare.”

For Somalis, one of the most valuable of the results they hope for from improving contact and communication with non-Somalis is that increased understanding on the part of non-Somalis about the experiences and culture of Somalis would foster better, more livable relations. Many also concede that Somalis do need a bit of extra encouragement and explanation when it comes to rules; “…to some degree Somalis don’t know about rules. Back home there was corruption...now we are newcomers here. Rules at home were dictatorial, now, here they know about the rules but they don’t take them as seriously.”

Many Somalis desirous of changing the attitudes and coercive practices of security, also recommended hiring a Somali security guard. One respondent wished that security would “cooperate with people” rather than always being “mean to the children and visitors, telling them to move and giving them a hard time. If they hired a Somali person maybe people could understand everyone. Sometimes when someone doesn’t understand the language they can seem rude. The Somali person could help by translating.”

Non-Somali residents were particularly adamant about the Somalis apparent disregard for the rules of the buildings. Some interpreted it as indicative of an inherent lack of consideration and were morally appalled. A particularly extreme position is provided by one resident who observed:

Nothing can be done. Even if you had Somalis involved and were talking to them, they don’t care. They don’t listen or care about anything except themselves. They will stand and talk in the elevator doorway while people are trying to get on. They don’t care. They have no sense of anyone but themselves.

Another more insightful but no less pessimistic view voiced by another resident is
that attempts to bridge the gap between the Somalis and other residents were unlikely to succeed due to the power of the moral denigration and character defamation of Somali refugees that had taken place all around them. As he put it, it wouldn’t even matter if Somalis were ‘owners’ rather than renters of the units. ‘...they would still be seen as the same - people see them as milking the system.’

Finally in July 1993 when Somalis took to the streets to protest their treatment at Dixon. The events which took place and the sentiments which underpinned them cannot in any simple way be explained by ‘overcrowding’. Nor can it be reduced to mere ‘cultural difference’. The broader context of this very preliminary ‘micro’ study of Dixon must be considered - a context of rising intolerance to outsiders. increased economic insecurity. distrust of government; a context of increased punitiveness and decreased compassion: a context in which the morally despicable figure of the undeserving fraudulent refugee was already well-familiar as was its companion figure of the despicable welfare cheat.

These events at Dixon. which developed over the first few years of the 1990s, provide a local window into the negative and coercive impact of exclusionary discourses at a particular place and time on a particular group. Somalis living at Dixon were constructed and acted upon as a problem. That some Somalis admit that they do tend to speak loudly. or that they don’t take some rules as seriously as others neither explains nor justifies the resort to coercive and confrontational tactics by the security personnel. The fact that overcrowding and the ‘transient’ nature of the many of the residents’ stays has resulted in circumstances and situations that pose, at the very least relatively minor inconveniences and irritations (such as increased people traffic resulting in slow and crowded elevator service, increased noise) and at worse increased financial costs of residency (hydro. water) neither explains nor justifies the racist and morally denigrating assumptions and stereotypes that fuelled the fire.

Somalis at Dixon were constructed and acted upon as a ‘problem’: by management. by owners, by residents and by security. Shortly after the demonstration in 1993. the City of Etobicoke held an emergency meeting and set up the ‘City of
Etobicoke/Dixon Task force" to deal with the 'problem' posed by Somalis, with a primary but not exclusive focus on the situation at Dixon. The Task force organized monthly 'community meetings' over a period of two years. Interestingly the 'community' was comprised of non-Somali residents and owners, security guards, parks and recreation, public health, political representatives and social services. At any given meeting there were only one or two Somalis in attendance. According to Maggie Redmonds, a community development officer for Metro Social Services who attended several of the meetings, the underlying point of view which was frequently made explicit was that the 'problem' inhered in Somalis.72

Several speakers sought to diffuse the panic about Somalis by addressing certain misconceptions and fears about the community. For example, the myth of widespread Somali welfare fraud was debunked by then commissioner of Metro Social Services Don Richmond and a representative from the department of Public Health sought to tone down fears about tuberculosis in the community.73 Nonetheless, the very fact that this Task Force was set up and these issues were addressed speaks to the degree to which the 'problem' of Somali refugees had become the focus and object of public and political concern.

In the very attempts to diffuse hostilities by addressing and dispensing with the various 'problems' thought to be posed by Somalis, the forum itself served to legitimate and reproduce the idea that Somalis were a 'problem'. This rather cynical view of the work of the Task Force is unfortunately supported by the fact that Somalis were not an active part of these 'community' meetings in which they were assumed to be a 'problem' and which sought to address the 'problem' (either through deconstructing it or through reproducing it) as well as the fact that the end result of the work of the Task Force was an "unsubstantive report with no real recommendations despite its apparent pro-Somali

72 Interview with Maggie Redmonds, former Community Development Officer. Metro Social Services. April 1999.

73 Ibid.
orientation." Redmonds describes the entire process as a public relations "smokescreen for not doing anything." 

Arguably, while the Task force did not succeed in producing any semblance of a coherent and proactive plan to address the issues raised, it did effectively reproduce the idea of the 'problem' of Somalis, supplemented by and exemplary of the broader 'problem' posed by undeserving and undesirable refugees and new immigrants in general.

### iii) Fanning the Flames in the Popular Print Media and the Enforcement Response

At the same time as the 1993 Federal Immigration 'intelligence' reports on Somali fraud and criminality were leaked, a similar allegation had been made in Toronto. As mentioned above, information had been received by the RCMP alleging that Somali employees of an Etobicoke Somali settlement services agency had misappropriated government funds allocated to the agency and that the funds were being sent back to Somalia to support the criminal activities of warlords in Somalia.

Two years later, in 1995, the very same allegations resurfaced in a letter written by a well-known Somali community member. This time, this allegation and others, made it into the popular press. In August 1995 the widely distributed magazine *Toronto Life* published a seven page article on the 'problem' of Somalis and what this problem should

---

74 In the year 2000, the same problems persist in one part of Dixon. According to Ali Mohamud, management of 3 of the 6 buildings has changed its security regime. They hired a new private security company whose officers more accurately represent the community, including Pakistani and Somali officers. In this part of Dixon, the problems of conflict and coercion have been largely taken care of. In the other part of Dixon, however, the same security company is in place, there is no minority representation amongst its officers, and conflict and coercion continues to be a serious problem. In December 1999, a complaint was lodged with the Ontario Human Rights Commission by a Somali resident of Dixon regarding the "racial insults, intimidation and the targeting of their children", of Somali residents by the Property Management company, including security. "Home Hassles" *NOW*. December 9-15, 1999

75 Ali Mohamud, Interview January 13, 2000; Maggie Redmonds, Interview. May 6, 1999
teach Canadians about their too generous, inefficient immigration system.76 The piece was written by Daniel Stoffman, already well-known in anti-immigration circles. Stoffman argues that the deleterious effects of an open and unselective immigration system must be considered. These effects, in his opinion, justify a more restrictive and selective immigration policy. In the August 1995 Toronto Life article, Stoffman is clearly intent on substantiating his position by providing a concrete example of the problems associated with an open immigration system. Once again, the ‘problem’ is posed by Somalis.

The impact of this article was significant. In addition to reinforcing and reproducing now familiar negative stereotypes of Somalis and adding to the shadow of suspicion that was already cast over the Somali community, the allegations contained in Stoffman’s article triggered another RCMP investigation into the very same allegations that had been made, and disposed of, two years earlier. It also precipitated at least one series of RCMP raids on the homes of suspected Somali ‘war criminals’ and several subsequent arrests and deportations. While Stoffman’s article outraged many, it clearly fanned the anti-immigrant/refugee flames and provided fuel for the enforcement fire already well underway.

a) “Dispatch From Dixon” and the (re)Construction of Somalis

In his article, “Dispatch from Dixon”, Stoffman begins by restating what had by then become rather commonplace and inflammatory allegations levelled at the Somali community living in Toronto. In bold type-face, Stoffman states that “many” Somalis “are cheating the welfare system” and that “some are probably war criminals”(my emphasis). These sweepingly general allegations which introduce the article provide the impetus and basis for what follows. Stoffman sets out to explain ‘how we created this mess’?

That there is ‘this mess’ and that it is evidenced by Somali welfare cheating and by the presence of Somali war criminals in Canada is uncritically accepted by Stoffman. Welfare fraud is but one of the allegations pursued by Stoffman in this article. He also

76 Daniel Stoffman, “Dispatch from Dixon” Toronto Life August 1995
attends to the criminality/fraud related allegations of corruption and cronyism amongst government officials in Toronto. of shady clan-based practices of ripping off the welfare system in order to send money back to Somali warlords, and of the presence of Somali war criminals in Canada. Stoffman's article is based in large part on interviews with a handful of sources without any evidence that they are either credible or reliable. Stoffman treats most of the allegations as 'fact'. These 'facts', made up largely of rumour and inflected with innuendo, are then treated as the supporting 'evidence' for Stoffman's own reflections on the more general issues of international migration, Canadian immigration and refugee policy, social welfare policy and administration. 77

Stoffman offers that "...anyone with dark skin who arrives at the border or an airport without documents and claims to be fleeing Somalia is allowed to apply for the status of a Geneva Convention Refugee." In fact individuals fleeing persecution are not 'allowed' to claim refugee status, they have the 'right' to do so under International Conventions and domestic legislation. The colour of their skin is not relevant to this right. Moreover, the fact that many do not have government issued travel documents is hardly surprising given the circumstances of their departure and, particularly in the case of Somalis, given the absence of a functioning civil authority in Somalia.

Further forging the link between Somali refugees and criminality, Stoffman ruminates provocatively about the husband of one of the Somalis whom he interviewed. He asks a series of provocative questions which lead the reader to think of her husband as complicit in the atrocities of the Barre government. He concludes this line of questioning by asking "Isn't this the sort of abuse that has brought our refugee determination system..."

77 Bob Swain, a 'sympathetic outsider', is another source for Stoffman. He is a biologist affiliated with the Canadian Baptist Ministries who worked with Somalis in Kenya and who is now actively involved in assisting in the resettlement of Somalis in Toronto. According to Stoffman, it is Swain who recounts stories of the misuse of public funds and the lack of accountability of public officials. Despite Stoffman's own admission that the tale is constructed out of rumours, they are each mentioned in turn. Significantly, a subsequent issue of Toronto Life carried an angry letter from Swain criticizing the irresponsible and misleading use by Stoffman of his interview. Toronto Life, October 1995
into disrepute?" Stoffman here transforms what is pure speculation on his part into a fabricated allegation of *bona fide* abuse. Even the layout of the article makes its own contribution. It carries bright red enlarged quotations which run across the top and bottom of each page: "The real beneficiaries of the world's most open refugee determination system are the Somali criminals of war"; "Some Somalis think of all governments as the enemy. They call welfare "shab" - meaning something for nothing".

At the close of the article, Stoffman makes reference to the stated wish of one of his Somali sources that the Somali's would "disappear", assimilate, adapt to the "Canadian way of life". Canadians are imagined as a homogeneous group, bound by common practices and sensibilities. Somalis are constructed as the problematic dark skinned alien 'other' whose practices and sensibilities are, by in large, criminal.

By 1995 it was beginning to feel like a campaign. Somali welfare cheats and warlords were routinely in the press. racism and coercion had tainted their home life: "[W]hen the Stoffman article hit the press, the Somali community was outraged."78 Members of the Somali community organized, held many meetings and strategized on how to respond to the unsubstantiated and discriminatory allegations contained in the article. Representatives of the community went to Metro Council and delivered a 'protest speech' to Councillors. Ali Mohamud remembers that this "...was a very difficult time for Somalis in Toronto", that Stoffman's article had been "...very, very damaging to the Somali community"; "The article was the starting point for all that followed. It triggered everything. Even some people, some anti-immigrant groups, duplicated and distributed the article, free of charge in front of the University Avenue Immigration offices."

**b) Enforcing the 'Problem'**

Shortly after the publication of Stoffman's article, the RCMP were back on the case again: the very same case based on the very same substantive allegations as in 1993. However, according to Mohamud, they readily conceded that their return was politically

---

78 Ali Mohamud. Interview January 13, 2000
motivated:

After one and a half years, they came back. The same RCMP officers as in 1993. They said that they were there because they needed to be able to say that they had talked to the manager of Dejinta Beesha, but that this was because of political pressure and that they knew that there was nothing. So they came, had a coffee, we talked about some other things and that was it. They wrote the same report that there was nothing wrong and that the allegations were false and unsubstantiated.\(^7\)

After Stoffman's article, the RCMP also contacted Peter Crosbie of the Family Services Association, one of Dejinta Beesha's trustees. According to Crosbie, there was no question that Stoffman's article was a central factor in their renewed interest in the Somali agency's financial operations. Indeed, they specifically referred to it, as well as to another 'community-source' in their explanation for their visit. Crosbie was first contacted by the RCMP by phone in early September 1995, was visited by an RCMP officer in late September and, "just when [he] thought it was all over", the officer came back to visit Crosbie in November, 1995. According to Crosbie, a member of the Somali community who was the primary source of the 'community-based' information attained by the RCMP, had written many letters to politicians and law enforcement authorities restating the allegations.\(^8\) The publication of these allegations in Stoffman's article in August 1995 tweaked the RCMP into action again. However, just as in 1993, no evidence was ever found of any wrong-doing.

Stoffman's article claimed legitimacy by giving prominent place to 'community-derived' information. It relied extensively on the comments and allegations of a member of the Somali community, indeed many of Stoffman's arguments and observations are initially raised by one of several Somali voices. Just as 'community-based' information had been the primary source of information for Stoffman's article so it was for the 1993 RCMP investigation, and, so it was for the 1995 RCMP investigation that followed the

\(^7\) Ibid.

\(^8\) Peter Crosbie, Director, Client Service Division, Family Service Association of Metropolitan Toronto, Interview, January 14, 2000
publication of Stoffman’s article. However, it appears that in all three cases, the information derived from the ‘community’ and upon which enforcement activity was justified was not only the same substantive information but it was also likely derived from the same community member; one Somali, many allegations, one article, two RCMP investigations, and the criminalization and denigration of an entire community.

These events highlight several important critical issues- the nature and use of community-derived information by authorities, particularly when that information has the potential of provoking an enforcement oriented response: the social construction of ‘problems’ (in this case ‘the fraudulent Somali war criminal’); the political dimension of enforcement responses; and the selective use of community-based information by authorities.

The enforcement fall-out against Somalis living in Toronto did not end there. Corruption and fraud were only two of the allegations given expression in Stoffman’s piece. The other particularly damning allegation made about Somalis was that ‘many were probably war criminals’. While this was not a new issue, it had been somewhat displaced by the increased attention being paid to the fraudulent character of Somalis. However, allegations of war criminality within the Somali community were indeed acted upon. One holiday weekend in 1995, the RCMP conducted several raids on the homes of Somalis in Etobicoke, who were suspected of being ‘senior members of a government which has committed war crimes or crimes against humanity.’

It will be recalled that this provision was included in the enforcement oriented Bill C-86 of 1992. According to Ali Mohamud, several people were arrested, including both former diplomats at Somali embassies and former military officers. It was also known in the community that a former chauffeur of a government official had been arrested, as was

---

81 Indeed, the problem had already been the topic of considerable contention and debate having received considerable public attention after CBC’s Fifth Estate aired a program on the subject several years prior.

82 Interview with Maggie Redmonds May 6, 1999
a former Minister of Education and an anti-female genital mutilation activist. Some of those arrested were subsequently deported, and some left voluntarily.

As was reviewed briefly in the last Chapter, the 1992 changes to the Immigration Act under Bill C-86 included a new provision to exclude individuals by virtue of the fact that they had occupied a ‘senior’ government position or had a senior government title in a government that had committed war crimes or crimes against humanity (s.19(1)). The provision is given effect by Order in Council which specifies the nations to which this provision will be applied. Somalia has been named and there is certainly no question that the prolonged and brutal dictatorship of Said Barre in Somalia qualifies.

However, the provision casts the exclusionary net widely, in large part as a result of systemic bias. It is the mere fact that a Somali was a ‘senior’ member of the Barre government that renders them inadmissible. It is assumed, through the application of Canadian standards, that the occupation of a ‘senior’ government position or the holding of a ‘senior’ government title necessarily reflects the degree of decision-making power and influence wielded by the person in question; their complicity is legally inferred and need not be proven. Moreover, the guidelines on the definition of ‘senior members’ and ‘senior officials’ are vague and explicitly open-ended.

Even where specified, the provision does not allow for the possibility that those in senior positions may have had no power over government policy at all and therefore no complicity. Moreover, while the provision is written to apply to ‘senior’ members of government, the fact of working for the government in any capacity has often been enough to raise suspicion and trigger an enforcement response. This heightens the impact of this provision on those Somalis who did not end up in refugee camps but managed to flee to Canada. Only Somali refugees with substantial resources could do this. They are

---

83 Ibid.

84 Interview with Ali Mohamud January 13, 2000

85 For the list of positions which inherently qualify as ‘senior’ see discussion of Bill C-86 in Chapter Four
generally well-educated and have considerable work experience. They are also therefore those refugees most likely to have occupied at least one, probably many, government positions under Barre’s 23 year rule. After all, in Somalia the government was the main employer. This provision has effectively criminalized most of the Somali refugees in Canada.

From chauffeurs to activists to military officers, all manner of government workers have been recast as being potentially complicit in the crimes of their employer. This range of positions formerly occupied by the Somalis who were arrested in the 1995 RCMP night-time raids illustrates the systemic bias of this provision in its implementation:

That piece of legislation was and still is unfair to the Somali community in Canada. It is well known that Said Barre ruled the country for 23 years. The government was the main employer, it was a socialist form of government, no private ownership...so. most people, when you finish university, you were placed in one of the government departments. You go from there, you work here, you work there, you become the head of the section. And since the salaries were so low, people were given some titles in order to get some extra benefit. So, to become a director was nothing. People compare it to being a director or a manager in Canada, but you can’t compare it. The practice was: you are a Minister and probably you don’t have any power. Or you could have been a secretary and you could have had lots and lots of power. It all depended on you relationship with government, with the ruling elite.

When people came here as refugees they told [the authorities] I was the director of that department, I was the member of the government. I did these jobs. And that was held against them. People said if you were in a decision-making position you are a war criminal. Not necessarily that you have committed any crimes against humanity, but the mere fact that you held that position excludes you; makes you a member of an inadmissible group. 86

After the Stoffman enhanced panic died down, so did RCMP enforcement activity in the Somali community. As observed by Mohamud, “...since that sweep and the

subsequent deportations. the RCMP and political interest has died down and we haven`t seen a lot of other deportations under that legislation”. Presumably this does not reflect a change in the seriousness with which the problem of war criminals is treated by enforcement authorities, nor does it indicate any change in the size of the problem or the success of enforcement efforts. It does, however, illustrate how, in an already hostile social and political context, the `problem’ of Somali war criminals was constructed and mobilized, in part by and through the popular media. And how the existence of this `problem’ triggered an enforcement oriented coercive response against a particular group of new immigrants and refugees.

iv) Proof of Identity: Prudence or Prejudice

In 1992, Bill C-86 effectively assured that the vast majority of Somali refugees in Canada would not be able to become permanent residents. The Immigration Act now requires that refugees (those determined to be ‘genuine’ by the Immigration and Refugee Board) provide ‘satisfactory’ (government-issued) identity documents in order to be granted landed status (permanent residency). The legislation also sanctioned the detention of refugee claimants without official papers. Two large groups of refugees already in Canada bore the brunt of the former requirement, Somalis and Afghans. Although as noted by Ali Mohamud, grouping Afghans and Somalis together has the effect of disguising the legislation’s particular impact on Somalis. Mohamud points out that the Somali community in Canada is much larger than the Afghani community and moreover "...sometimes Afghans can get documents. There is a government there. It’s easier.”87 In contrast, there is no functioning government in Somalia; "...there is no civil authority. Everybody knows that there is no government in Somalia. There is a civil war. You can’t get any documents. You can’t get a passport, you can’t get anything."88

The denial of permanent residency status is indeed a severe blow to those already

87ibid..

88ibid..
displaced and disempowered. Without it, refugees who have been found to be ‘genuine’. cannot travel outside Canada, sponsor relatives or attend university. They are ineligible for many jobs and training programs and cannot obtain a bank loan. The inability to sponsor family members is a particularly painful consequence for many Somalis who had to leave behind close family members when they fled their country. As put by the Canadian Council for Refugees, “Lack of identity documents for these refugees means that they cannot get on with their lives. It means they cannot heal. It means the persecutors have in a way ‘won’.”

The anger felt by Somalis left in legal limbo by this legislation was further heightened by the fact that just as Bill C-86 took effect, the federal government had taken measures to facilitate the acceptance of 26,000 refugees from the former Yugoslavia. Critics, while supportive of helping these refugees, were quick to point out that no such measures had been initiated to help Somali refugees and that furthermore, the newly adopted safe third country provision would have precisely the opposite effect. As put by one refugee lawyer during the Legislative Committee hearings on Bill C-86, while advocates supported the move to help refugees from the former Yugoslavia.

...we do not understand why it is in the Horn of Africa in situations like this. where frankly, many, many more people are being killed. many more


90 Ibid.

91 The Refugee Lawyers Association strongly condemned the Safe Third Country provisions on the ground that “...this provision discriminated against black, brown and yellow people. against people from the developing world. The bottom line is that most people who can get to Canada on connecting flights are people who are coming from European countries. If you come from Asia or from Africa or from South America or Central America. you simply cannot do that...We’re very concerned for many reasons...when you look at the statistics of the IRB, you’ll see that since 1989 the majority of people who have received protection have been from developing countries, mainly from Somalia and Sri Lanka....This bill is being considered precisely at a time when the situation for these people is most desperate.” (Quoted in Jakubowski 1997:86-87)
Mohamud observed simply, "...if we had been from the former Yugoslavia, we would have been landed a long time ago."  

The official response to criticisms of the unequal treatment of different groups of refugees by the government served to further malign and anger Somalis. Bernard Valcourt, then Minister of Immigration, explained that the Somalis situation was less compelling owing to the fact that they were 'nomads' who didn't want to come to Canada anyway. This remark sparked a protest at Queen's Park in Toronto by Somalis angry at this dismissive and offensive official response to their very desperate circumstances.

Ostensibly, the difficult and differential impact of this requirement on Somali refugees was officially acknowledged in 1997, when the government announced measures to deal with undocumented refugees. The Minister stated that many refugees from Somalia and Afghanistan "have been unable to obtain proper documentation due to sustained civil war and lack of an effective government authority to issue identity documents." In response to this situation the government introduced new regulations which would allow Convention refugees, from specified countries, to become permanent residents five years after a positive refugee board decision.

Many then argued that the waiting period was unfair, given that many of the refugees had already waited up to three years to have their claim heard, the government was, in the end, not persuaded. The persuasiveness of the humanitarian argument was

---

92 Quoted in Jakubowski 1997:86

93 Ali Mohamud. Interview January 13, 2000

94 Toronto Star. August 16, 1992

95 "Lucienne Robillard Announces the Introduction of the Undocumented Convention Refugees in Canada Class" News Release, Ottawa: January, 1995
eclipsed by law and order concerns and by corollary notions of deservedness. The Minister's justification for imposing the five year waiting period is quoted here at length:

I have carefully reviewed all the comments received during the pre-publication period. I understand and share the humanitarian concerns expressed by certain organizations, which argued that we should find a more generous solution to the difficult situation that Somali and Afghan refugees are experiencing in Canada. I certainly would respond differently if my only concern was the specific circumstances of the affected communities. As the Minister responsible for immigration to Canada, I must ensure that every effort is made to discern the background and character of applicants for permanent residence. At the same time, the asylum system must not be abused by those who may choose to conceal their identity as Canada continues to be generous to those who really deserve protection.  

The Minister a year later made the same point even more explicitly. "Because they have no ID, we will not grant these people permanent resident status until they have had time to demonstrate respect for the laws of Canada and for us to detect those who may be guilty of crimes against humanity or acts of terrorism...The message is clear -- fraud will not be tolerated."

As mentioned earlier, the method of 'detection' employed by authorities relies in a central way upon the gathering community-based information. As explained by Brian Grant, Acting Director General of the Department of Citizenship and Immigration in 1998, this is an important dimension of Immigration enforcement practices particularly when identity is an issue:

...we have tried to build flexibility into the system. precisely because you often don't know who they are. They don't have a document. They've come from outside North America so you don't have linkages. The police system may not be reputable in the country they've come from. They will

---


gravitate towards their community and often that happens is that people in the community might recognize them and come forward. You’ll often get tip-offs like oh, so-and-so, we know him and he was involved in this. That will often give us the lead we need to follow up on security threats or perhaps criminality.  

In response to this statement made in 1998, another member of the Standing Committee pointed out that this method of gathering information “creates its own problems too, particularly in the Somali community where one group doesn’t like the other group” to which Grant responded “yes, you have to sort through all that.”

While community-based information is treated seriously enough by authorities in the context of exclusionary enforcement activities, to the degree that it may justify extreme, state sanctioned coercive responses, it appears that when community-based information is offered as a way of confirming identity for the purpose of inclusion, authorities are less willing to act on it. In 1996, prior to the announcement of the 5 year waiting period, representatives of the Somali community announced a legal challenge of the identity document requirement on the grounds that it discriminates against their community. Representatives of the Somali community suggested that in the absence of documents and given the relative impossibility of acquiring them, that the community’s “elders and clergy would verify the authenticity of the claims.” They stressed that they share “...the federal government’s concern that sworn affidavits could be abused to allow criminals and imposters enter the country” and that clearly Somalis “...have no desire to live alongside law-breakers.” As evidence of their status of good ‘deserving’ citizens and of their credibility, they observed that “…the community has already identified 6 members of the former Said Barre regime to the RCMP and CSIS.”

The difficulty of having inclusionary identity related community-based

---

98 Brian Grant, Standing Committee Minutes, Thursday February 19, 1998:10-11
99 Ibid., p.11
100 Globe and Mail “Refugees to Challenge Immigration Act, Somali Group Alleges Discrimination” February, 6. 1996
information acted upon by authorities stands in stark contrast to the willingness of authorities to act upon similarly identity-related community-based information when that information is negative and exclusionary. It also highlights the different uses of dominant discourses. Criminality and fraud discourses produce both the undeserving and the deserving new immigrant and refugee. In the contemporary discursive climate, new immigrants and refugees frequently seek to (re)present themselves as law-abiding, and otherwise deserving; arguably this is one of the reasons that enforcement oriented efforts have been effective in soliciting ‘tips’ from the community.

In this instance, fraud and criminality concerns converged and were mobilized in direct opposition to humanitarian ones in the official justification for the five year waiting period. Moreover, resort to criminality/danger and security discourses legitimized the use of various coercive and invasive means to ‘contain the threat’. So, for example, in addition to the rather punitive consequences facing Somalis without identity documents who must wait years for family reunification, they are often ‘...forced to undergo expensive, time consuming and invasive DNA testing to prove their family ties.”

viii) The Criminalization of Khat

There is little question that the impact of the convergence of fraud and criminality discourses over the enforcement oriented decade of the 1990s was felt particularly sharply by Somalis in Canada. They have been the subject of sustained identity ‘reconstruction’ since they arrived in Canada as refugees. This reconstruction from that of legally-confirmed) genuine victims of persecution deserving of protection to that of (socially.

---

101 Canadian Council for Refugees, “CCR Calls on Canadians to Respond Warmly to Refugees” May 6, 1999. According to Ali Mohamud of Definta Beesha, (interview, January 13, 2000) the announcement in 1999 made by the new Minister of Immigration, Eleanor Caplan, that the 5 year waiting period was to be reduced to 3 years in recognition of the difficulties it poses for so many refugees from Somalia and Afghanistan has not been translated into practice. Indeed, he reported that in January 2000, nobody from the Somali community had yet been landed under the 3 year rule. In fact, Immigration officials had reported to members of the Somali community in January that they had received no official directive regarding the new policy.
politically and publicly constituted) undeserving fraudulent criminals. entailed concrete and coercive consequences for Somalis.

There is one final entry in this (re)constitutional tale. The fact that many Somali refugees in Canada are living in legal limbo with no affirmed or protected citizenship status makes them particularly vulnerable to deportation for reasons of criminality. This fact, coupled with the fact that Somalis have been increasingly criminalized in Canada over the 1990s, means that the spectre of deportation looms particularly large for the community.

This reality was further entrenched in 1997 when the Canadian government criminalized the plant Khat under the new *Controlled Drug and Substances Act* (CDSA). The CDSA was an enforcement oriented and prohibitionist piece of legislation. It consolidated Canadian drug policy in accordance with international obligations. It repealed the *Narcotic Control Act*, and Parts 3 and 4 of the *Food and Drugs Act*. The chemical properties that justify the inclusion of Khat under Schedules 3 and 4 of the CDSA is that it contains cathinone, a ‘psychoactive ingredient’ and sometimes d-amphetamine. However, illegal drugs, like illegal activities, are not simply objective unchanging facts; they do not exist in nature.

Khat is a plant with stimulant properties that is grown in East Africa and the Arabian peninsula and which is known to be popular with some members of the Somali community. Chewing Khat has long occupied a central place in the cultural tradition of male Somalis (a fact, which was duly noted in Stoffman’s 1995 article). As explained by

\[\text{---\textsuperscript{102} The Controlled Drug and Substances Act, Ch.19. Canadian Statutes of Canada, 1996}\]

\[\text{---\textsuperscript{103} Diane Riley, “Drugs and Drug Policy in Canada: A Brief Review and Commentary”, Canadian Foundation for Drug Policy, November 1998}\]

\[\text{---\textsuperscript{104} Diane Riley, “Drugs and Drug Policy in Canada”, (A study prepared for Senator Pierre Claude Nolin as a background document for his June 1999 motion to have Canada’s Senate conduct a thorough review of Canadian drug law and policy). Canadian Foundation for Drug Policy, 1999}\]

\[\text{---\textsuperscript{105} Ibid.}\]
Farah Khayre: "Khat in the Somali culture has traditionally been used socially, much like coffee in the western culture. It has no criminality associated with it." While Ali Mohamud concedes that it may be the cause of some "family conflict" due to its gendered consumption, he notes that the degree of conflict associated with Khat does not compare with that of alcohol, for example: "...if you eat some, you might want to eat some more...but it doesn’t make you want to kill or fight; it’s a harmless pastime."

The United States criminalized Khat not long before Canada did. The American inclusion of Khat as a Schedule I narcotic (the most restricted category), alongside heroin and cocaine, derives from the exposure to the plant which US marines had when they were posted in Somalia as a peace keeping force in the early 1990s and which was subsequently brought to the attention of the drug enforcement authorities. The United States subsequently placed Khat on its list of prescribed substances and urged other governments to do the same.

Canada followed the American lead. Prior to 1997, the use of Khat was legal in Canada while bringing it into the country was not. It was regulated as an ‘alien plant’ by the Canadian Food and Drug Inspection Agency and could be seized by Canada Customs at the border. However, Khat had become the focus of official attention before 1997. In the midst of the panic surrounding Somali refugees in Canada, Canada Customs had stepped up their enforcement of Khat related import violations.

The CDSA creates six major offences: possession, trafficking, cultivation, import.

---


107 Ali Mohamud. Interview, January 13, 2000


export and prescription shopping. Since 1997, the possession of Khat is a hybrid offence, that is it may be processed summarily or by indictment. For example, if possession is dealt with as a summary offence, the punishment is up to 6 months in prison and a $1,000 fine for the first offence. 12 months in prison and a 2,000 fine for subsequent offences. If it is proceeded with by way of indictment, the maximum punishment is 7 years in prison.\textsuperscript{111}

The consequences of a conviction for possession of Khat for an individual who does not have permanent residency status in Canada are extremely serious. Under s.4(2.1) of the Act, a person found to be a Convention refugee loses their “right to remain in Canada” if they have been convicted of an offence under any Act of Parliament for which a term of imprisonment of: a) more than 6 months has been imposed; or b) five years or more may be imposed. For the majority of Somalis in Canada, therefore, a conviction for the possession of Khat has the real potential to trigger deportation; because of the administratively imposed delays on their access to Canadian permanent residency, they have no secure status to protect them from removal.

Another significant dimension of the CDSA is its strongly ‘prohibitionist’ orientation and resultant emphasis on empowering enforcement activity. The CDSA sanctioned increased enforcement powers for those engaged in the ‘war against drugs.’ Among other provisions, the Act extended the powers of search and seizure. In the 1995 Legal and Constitutional Affairs Standing Committee Hearings on the Bill which created the CDSA (Bill C-8), Robert Kellerman, a member of the Law Union of Ontario, forcefully criticized the enforcement orientation of the Bill from a legal, rights-based perspective. Kellerman argued that the enforcement of drug laws has resulted in the court sanctioned “...erosion of many of our civil liberties.” despite the protection of privacy rights afforded by the Charter. He noted that “...an overwhelming number of searches

\textsuperscript{111} Diane Riley, “Drugs and Drug Policy in Canada” 1999
take place under our drug laws."\textsuperscript{112}

The clauses dealing with seizure or forfeiture of property under the Act are broad. As observed by Kellerman: "...This clause has the danger [that]...property could be taken away from people who are totally innocent." He explained that: "The idea is that the property of drug dealers can be seized. Yet it is so broadly defined that it includes even the seizing of property for the simple offence of possession of marijuana."\textsuperscript{113}

Since Somalis make up the vast majority of Khat consumers in Canada, the consequences of criminalization were for the most part limited to them. And the consequences were severe; involving the mobilization and application of both criminal justice and immigration enforcement systems and sanctions. Soon after the criminalization of Khat, in the spring of 1998, York region police carried out numerous raids on Somali residences at Dixon on the grounds of suspected Khat related offences. Armed with the increased powers of search and seizure described above, officers did both. They confiscated substantial amounts of gold and jewellery without providing receipts and according to reports they also damaged other items of Somalis' property in the course of carrying out the raids;

Police from York region came to Dixon Road, invaded people's homes, confiscated jewellery and money without explaining anything, in the name of drugs. in the name of Khat. Khat is a tradition that Somalis chew and they made it one of the controlled substances. They know that Somalis chew Khat. And you cannot equate Khat with other narcotics. So it was one way of criminalizing Somalis.\textsuperscript{114}

According to members of the Somali community, the police engaged in unprofessional and downright nasty conduct during these raids:"[T]hey wore trench coats and they had their pistols and they would come into our homes and destroy everything:

\textsuperscript{112} Robert Kellerman. Law Union of Ontario, Submission to Standing Committee on Legal and Constitutional Affairs", transcripts, December 13, 1995

\textsuperscript{113} Ibid.

\textsuperscript{114} Ali Mohamud. Interview, January 13, 2000
break personal property and destroy computers, break everything...”

What is both particularly troubling and revealing is there had been no prior effort to communicate the changes in the law and in enforcement priorities to the Somali community. In a background paper prepared for Senator Pierre Claude Nolin’s June 1999 motion to have Canada’s Senate conduct a thorough review of Canadian drug law and policy, Diane Riley of the Canadian Foundation for Drug Policy made specific mention of this rather shocking fact. Riley observes that the inclusion of new drug offences under the CDSA ensures “...that more Canadians than ever...would be burdened with a criminal record for simple possession (this is already occurring with respect to Khat, especially among the Somali community who were not even informed of the criminalization of activities related to a previously legal substance...” (my emphasis). According to Farah Khayre of Midyanta,

The law was passed quietly, not even fully debated in Parliament, no community information was sought, and [there was] no outreach to the community for information about this law. We just started getting calls from people who had been arrested or their homes had been broken into.

Kellerman’s dire predictions about the abuses associated with the strict enforcement of drug laws were indeed accurate. According to reports, homes were searched and property seized of Somalis who had never even used Khat, let alone sold it.

Outraged once again, the Somali community organized a large community meeting to decide how to respond to the criminalization of Khat and the severe enforcement response that had been executed in the Somali community; “[M]embers of the Somali community are in shock over a sudden police crackdown against the substance Khat. with attendant civil rights abuses, and are seeking to have police searches of their

115Ibid..

116Farah Khayre, Midaynnt, Association of Somali Service Agencies. interview quoted in "Khat: What the Hell’s That"

117Ali Mohamud, Interview, January 13, 2000
homes reigned in, and the legislation that banned the substance last year repealed.”

The number of people convicted under the CDSA for Khat related offences is in the hundreds, and the number of arrests much higher. For Somalis living in limbo in Canada, a conviction for Khat means “...that’s the end of it, you are gone...”: “Hundreds and hundreds of Somalis have been wrongfully convicted of drug charges and other minor things and they were refused to be landed for the reason of criminality.” Indeed, even in the absence of actual criminal convictions, the prohibition of Khat has served to further criminalize Somali refugees in the public eye. And unquestionably it has contributed in a powerful and blatant way to both the entrenchment and reproduction of the undeservedness of Somalis and the coercive crime control response which that undeservedness elicits today.

These consequences had been anticipated; the government had been warned. During the course of the Standing Committee on Legal and Constitutional Affairs hearings on Bill C-8 (CDSA), the consequences of criminalizing Khat were raised as a particular concern. Perry Kendall, President and CEO of the Addiction Research Foundation (ARF), warned strenuously of the dangers of criminalizing drugs and intensifying enforcement. From a public health harm reduction perspective, Kendall observed that criminalization and tough enforcement has been associated with the magnification of the potential health harms posed by certain substances, and that this was a pattern that the criminalization of Khat could provoke:

Aggressive enforcement has been associated with public health harms, including the phenomenon of illicit substances becoming purer, more potent and cheaper in recent years. Some suggest that the prohibitionist policies south of the border have increased the size of illicit markets, the potency of drugs and the profit from the trade.

We are concerned that, in a tiny way, the same pattern could happen to Khat, a drug which in plant form is popular in East Africa and the Arabian peninsula and is used by a considerable percentage of some immigrant

118"Khat: What the Hell’s That"

groups in Canada.

Criminalizing this drug as this bill would - and the drug is currently only available in the form of a vegetable - could have the unintentional consequence of encouraging sellers and markets to refine it into a more potent power whose effects would be potentially more severe than the current use of the vegetable.\textsuperscript{120}

Arguably, this ‘public harm’ argument against criminalization held out the most promise to be persuasive in the enforcement oriented, prohibitionist context of the bill. However, Kendall also raised concerns relating to the discriminatory and inequitable enforcement of drug laws in general and of the criminalization of Khat in particular: “We are concerned that prohibiting Khat would create a new form of a dangerous drug and criminalize a whole community in this country.”\textsuperscript{121}

The warnings were not heeded. The concerns raised regarding the consequences of the criminalization of Khat were well-founded. It did criminalize an entire community. it did facilitate the inequitable application of drug laws. it did facilitate abuses of the civil rights of Somalis. and it did bring in tow extremely severe and coercive consequences for Somalis, including, police raids, searches, seizures, arrests and removals. It also had wider implications, as observed by Eugene Oscapella of the Ontario Bar Association:

The war on drugs allows us to dress our racism and xenophobia in less obvious trappings. In North America, our early drug laws were to a significant degree premised on the vilification of immigrants and people whose skin colour or ethnic culture did not make the grade... This [still happens]. In Canada for example, in 1996, criminalized a stimulant called ‘Khat.’ Khat is a substance used by some people of African origin. It is not used by white Canadians of European descent to any appreciable extent, if at all...Why prohibit these recent immigrants to Canada, these people of another culture, another skin colour and another continent - from using a

\textsuperscript{120}Dr. Perry Kendall, President and CEO of the Addiction Research Foundation, submission to the \textit{Standing Committee on Legal and Constitutional Affairs} hearings on Bill C-8 (CDSA), evidence, Dec.13, 1995

\textsuperscript{121}ibid.. emphasis added.
viii) Conclusion

Genuine refugees who are found to be ‘truly deserving’ are still for the most part rhetorically and legally welcomed by Canada. However the presumption has taken hold in the 1990s that most refugee claimants are not genuine victims, but rather are fraudulent unscrupulous and dangerous opportunists and criminals. Whereas the undesirable and undeserving ‘other’ of the past were variously excluded through the mobilization of moralistic and racist discourses of national purity or through the political justification of national security defined more narrowly as a political threat, the contemporary construction of the undesirable and undeserving other is excluded through the mobilization of a more expanded version of national security which now encompasses not only the traditional ‘political’ threats posed by subversion and espionage and national terrorism but also international terrorism, organized crime and serious criminals who pose a threat to public safety.

Relatedly, as has been argued here, immigration exclusions in the 1990s have been increasingly governed by fraud and criminality/danger discourses, a development which has also taken place in the context of domestic social welfare provision. Moreover, as this chapter has demonstrated these developments are not merely abstract shifts but entail concrete and coercive consequences those subject to their operations. These dimensions are not readily apparent in official law and policy. They become more apparent when certain groups become the focus of national and local political and public panics. 123


Certainly other groups have been variously constructed and acted upon at different times, and in different locations, including: Eastern European ‘gypsies’, Chinese triads, Jamaican drug dealers, Middle-Eastern terrorists, Nigerian frauds, Eastern European strippers.
Chapter Six

Discretion, ‘Danger to the Public’ and ‘Just Desserts’:
The Legal Exclusion of Permanent Residents in the 1990s

I) Introduction

Chapter 5 mapped the emergence, mobilization and conflation of criminality/danger and fraud discourses in the governance of Canadian exclusionary immigration law and policy in the late 1980s and early 1990s. These developments were then examined in terms of their impact on the construction and regulation of the Somali community. In this period, criminality, as an exclusionary definitional category, was steadily redefined and expanded as representing a threat to ‘national security’. Parallel to this, over the decade of the 1990s, the dominant understanding of national security has been extended and modified to encompass various forms of criminality. Evidence of this discursive development is provided by the expansion in the 1990s of the official mandate of the Canadian Security Intelligence Service (CSIS). CSIS’ concerns over ‘threats to national security’ had previously focused on threats posed by terrorism (“...the planning or use of politically-motivated serious violence”) and espionage. (“...undeclared foreign intelligence activity in Canada and detrimental to the interests of Canada”).¹ In the 1990s, organized ‘transnational’ criminality was incorporated into the CSIS mandate. This was no small development.

Criminality has always been, to varying degrees, an effective and adaptable supplementary definitional category for the construction and exclusion of 'undesirable' immigrants and refugees. In the 1990s it has emerged as the dominant discursive rationale for exclusionary immigration policy operating in and indeed fortified by the definitional shadow of the spectre of ‘security’.

This chapter focuses upon the legal construction and regulation of the problem of the ‘criminal immigrant’ and the nature and use of ministerial discretionary power in this

¹The Canadian Security Intelligence Service Website, ‘The CSIS Mandate’
regulation. It begins with a discussion of discretion and immigration exclusions, followed by a more detailed discussion of the legal construction and regulation of the ‘criminal immigrant’ through a tough and major piece of legislation. *Bill C-44* in 1995. The social and political context, the legal content and the implications of this reform are examined in some detail with particular attention being paid to the place and use of discretionary power preserved by and through the Bill. The specific question of discretion is front and centre in this legal reform which sanctioned the deportation without appeal of even permanent residents who are deemed by the Minister to represent a ‘danger to the public’. It will be argued here that this radical extension of Ministerial discretion to exclude ‘serious’ criminals deemed likely to pose a ‘danger to the public’ under *Bill C-44* in 1995 can be read as the political and legal manifestation of the powerful shift in the governance of exclusionary Canadian immigration law and policy; a shift which has redefined, expanded and merged ‘criminality’ and ‘security’ concerns and which has brought both to the fore of exclusionary Canadian immigration policy.

II) Discretionary Power and Exclusion

As emphasized in Chapter Two, the context of discretionary power must be taken seriously. Within Canadian immigration law and policy, discretionary power has always been preserved and employed in the historically-specific work of identifying and excluding undesirable non-citizens and producing and reproducing the desirable Canadian citizen. As such, discretion is better regarded as a productive activity rather than a residual and latent ‘space’; discretion picks and chooses, assesses and evaluates, judges and weighs, denies and grants, bestows and withdraws, includes and excludes, produces and negates; in short, it governs. Discretionary power affirms and enforces historically specific constructions of desirability.

---

*An Act to Amend the Immigration Act and the Citizenship Act and to make a consequential amendment to the Customs Act, S.C.1995, c.15* [hereafter *Bill C-44*]. Royal Assent received 15 June 1995; came into force 10 July 1995.
Discretion is clearly a central mechanism in the normative and fundamentally moral project of constituting ‘desirability’ and negating ‘undesirability’. Yet under a liberal regime of governance discretion is officially constructed and regulated as disinterested, objective and professional evaluation and calculation. If it is employed unjustly or unfairly this is attributed to inadequate legal or legalistic regulation leading to ‘arbitrariness’ and the undermining of procedural fairness and objectivity. Chapter 3 demonstrated that in the context of immigration exclusions, discretionary power has historically always been employed in a particular kind of exclusionary ‘national service’. Facilitated and inflected by and through the operation of dominant discourses, discretionary power is a central technology in the production and enforcement of historically specific conceptions of the desirable and undesirable, the deserving and undeserving. It is as well a central mechanism which both facilitates and justifies the application of coercive ‘sovereign’ state power against ‘undesirable’ non-citizens.

Exclusionary discretionary power thus facilitates the enforcement of different, historically specific discursive constructions of desirability and undesirability. In the present day these are framed primarily through criminality and danger discourses. This view of discretion stands in contrast to the conventional welfarist liberal view of discretion as a compassionate provision for the exercise of mercy. It challenges the liberal legal view of discretion as the unruly corollary of law which facilitates the attainment of individual justice as long as it is adequately regulated to protect against ‘arbitrariness’ and tyranny.

While it is true that exclusionary discretionary powers are and have always been to varying degrees inflected by racist discourses and have often been applied in racist and otherwise discriminatory ways, the analytical approach followed here resists the analytical temptation to explain either discretion (or social constructions of ‘undesirability’) by reference to any single, universal axis of oppression or to any single governmental

---

5Dessert’ discourses are predominant in the context of inclusionary, humanitarian regimes of immigration decision-making. Discourses related to exclusions revolve around dominant notions of (un)desirability.
preoccupation. While not disputing the central influence that racist discourses have had in the development and application of exclusionary Canadian immigration law and policy, it is important to preserve analytical space for the consideration of the myriad of different and often conflicting discourses which inflect discretionary power in this field. Dominant constructions of criminality which underpin and justify contemporary allocations and uses of discretionary power in the exclusion of undesirables are not in any simple or straightforward way necessarily or exclusively reducible to race.

III) Exclusion, Discretionary Power and *Bill C-44*: A Detailed Interrogation

In July 1995, *Bill C-44* became law in Canada. Among other changes, it amended the deportation appeal provisions of the *Act* in the form of s.70(5) and in so doing this reform represents the most radical extension of wide-ranging Ministerial discretionary power to exclude since the 1952 *Act*. In this present case, Ministerial discretion is specifically aimed at ‘criminals’, in particular those who are deemed likely to represent a ‘danger to the public’. This amendment provides a powerful example of the contemporary dominance of criminality and danger discourses in immigration exclusions and of the way in which discretionary power works to facilitate the translation of dominant social concerns into concrete and coercive exclusionary immigration law and policy. It also represents a clear example of the ways in which systemic forms of discrimination in the context of criminal justice enforcement intersect with and influence exclusionary immigration law and policy.

This amendment has elicited much controversy. The new provision, s.70(5), permits the deportation without appeal of landed immigrants (permanent residents) deemed by the Minister to represent a ‘danger to the public’ regardless of how long they have resided in Canada. It provides for the deportation of people who may have lived in Canada for many years but who, for one reason or another, did not acquire Canadian citizenship. Also contentious and analytically significant is the possibility preserved by the amendment of the forced removal even of refugees.
This chapter begins with a brief discussion of the contemporary context, content and debates surrounding s.70(5) as contained in *Bill C-44*. It then examines the different ways in which scholars have sought to make sense of this recent legislative development.

i) The ‘Just Desserts’ Bill: Context

In April, 1994 Georgina Leimonis, a 22 year old white woman was shot and killed by three black men in ‘Just Desserts’, a small cafe in an affluent neighbourhood in downtown Toronto. In July of the same year, Police Constable Todd Baylis (also white) was shot and killed by black, Jamaican-born Clinton Gayle. The massive and sensationalist media coverage of these two events tended to gloss over the facts that: a) only one of the black men, Oneil Grant, involved in the Just Desserts killing was in fact Jamaican born, the other two men involved were Trinidadian and Canadian: and b) both Grant and Gayle had come to Canada as children.  

These murders, coupled with already heightened tensions between the black community and the police, triggered a massive public panic around the issues of race, crime and immigration, a panic which the government seized upon and responded to swiftly, decisively and on several different fronts. In brief, *Bill C-44* was first tabled on June 15, 1994 and was passed into law in July, 1995. The ‘danger to the public’ provision (s.70(5)) was not the only reform contained in *Bill C-44*: it was one of several which aimed to streamline enforcement through, for example, the expansion of the circumstances under which deportation orders are issued, the extension of powers of arrest and seizure of documents and increased powers to deal with multiple or fraudulent refugee claims.  

---

4 Julian Falconer and Carmen Ellis “Colour Profiling: The Ultimate Just Desserts” (Draft) 1998:16

On July 7, 1993, Sergio Marchi, Minister of Citizenship and Immigration, announced the reorientation of Canada's removal policies; a reorientation termed the 'criminals first' detention and deportation priority. Soon after, the Minister of Immigration and Citizenship also announced the creation of the Joint Immigration RCMP Taskforce to assist, among other responsibilities, in the tracking down and arrest of non-citizens with criminal records and who are wanted for removal. Also in 1994, the Department of Citizenship and Immigration announced the creation of an Organized Crime Section. Certainly, the expansion of the mandate of CSIS to encompass certain varieties of 'organized' criminality also in the mid-1990s is of central analytical importance as well. These developments will be discussed in more detail in Chapter 7. For now, the important point is that Bill C-44 was the first of a flurry of government initiatives following the 'Just Desserts' murder that aimed to 'get tough' on criminal immigrants - hence its informal reference: the 'Just Desserts Bill'.

The current linkage between immigration and criminality evidenced by Bill C-44 must be understood in its historical context; a context that long predates 'Just Desserts'. Todd Baylis and Bill C-44. As argued in this thesis, the historical development of Canadian immigration law and policy consistently reveals a central and guiding preoccupation with the exclusion of undesirables and moreover, that broad tracts of discretionary power have always been preserved in order to facilitate the exclusion and/or expulsion of those deemed undesirable. It is also the case, as described in Chapter Three, that the seeds of the contemporary governing rationale of protecting the security of the state and the public from criminals were sown long before 1995.

Moreover, the force of the current political preoccupation with getting tough on 'criminal immigrants' reflects a more general contemporary concern with crime and criminality evident in most areas of public policy. The current social and political climate

---

"More recently as the trial of the accused has slowly and painfully moved through the criminal justice system, slowed by allegations of racism and illegalities by the defence, the case has been referred to as the 'justice deserted' case. "Just Deserts Case a Mess. Judge Reveals" The Globe and Mail, Nov. 11, 1998"
is in large part fuelled by populist concerns about criminality, danger and victims resulting in the adoption of increasingly punitive measures in a wide range of governmental spheres.

In the 1990s, "get tough" politics have become everyday politics as crime and justice receive unprecedented levels of media coverage. Popular calls for draconian measures against criminals fuel a climate of punitiveness that politicians find irresistible for fear of appearing soft on crime.  

Immigration law and policy has not been immune to this. Indeed, as documented in Chapters 4 and 5, the conceptual slippage between criminals and immigrants is more pronounced today than it has ever been. "Criminal immigrants" have become Canada's public enemy number one. Exclusionary Canadian immigration law and policy has become an instrument of crime control engaged in the work of minimizing the risk of the victimization of "the state", "the nation" and "the public". The association between criminality and immigration exclusions is not, in and of itself, a new development. However, the emergence of criminality and danger discourses as the guiding rationale in the governance of exclusionary immigration law, policy and practices is indeed historically unprecedented.

Understood in this wider context, Bill C-44, while extreme and troubling, is not surprising. And Bill C-44 has indeed troubled a lot of people. Criticism has been largely focussed on the radical extension of Ministerial discretion sanctioned by s.70(5), namely the Minister's power to exclude permanent residents on "criminal" grounds, without appeal, regardless of how long they have resided in Canada. As put by lawyer David Matas. "...if somebody came here at the age of 5...they didn't come here as a criminal. They came here as a child."  

__________________________

7Carolyn Strange. Qualities of Mercy 1996:14

8Allan Thompson, "Falling Through the Cracks", The Gazette. July 2. 1994
ii) Legislative Content of Bill C-44

This then was the context of the presentation of Bill-C-44 to the House in June 1994 by then Immigration Minister Sergio Marchi. As mentioned, the effect of s.70(5) is to remove the right of a permanent resident to appeal a deportation order to the Immigration Appeal Division (IAD) under s.70 in cases where the individual in question a) has been convicted of a crime for which the maximum available penalty is ten years or more and b) has been deemed by the Minister or her delegate to represent a "danger to the public". There are now two categories of non-citizen 'criminals' who are excluded from the normal deportation appeals process: those deemed a 'security' threat to the nation and those deemed a 'danger to the public'. The former deals with what are conventionally understood as 'political' crimes and the latter with so-called 'true crimes' - although the distinction between the two has been increasingly blurred. The text of the provision relating to the latter category reads as follows:

No appeal may be made to the Appeal Division by a person described in subsection (1) or paragraph 2(a) or (b) against whom a deportation order or conditional deportation order is made where the Minister is of the opinion that the person constitutes a danger to the public in Canada and the person has been determined by an adjudicator to be c) a person described in paragraph 27 (1)(d) who has been convicted of an offence under any Act of Parliament for which a term of imprisonment of ten years or more may be imposed. (emphasis added)

iii) Administrative Guidelines, Appeals and the Question of Discretion

The Department of Citizenship and Immigration issued very sparse guidelines to assist in the forming of a 'danger opinion' by the Minister's delegates. These guidelines echo those which guide adjudicators in carrying out their regular 30 day review of

---

9 However, as will be discussed, those deemed to be 'security threats' are afforded significantly more legal protections than those deemed to constitute a 'danger to the public'. The substantive significance of this difference will be taken up in Chapter 7.

10 Immigration Act 1976-77, c.52, s.1, s.70(5)
detention decisions. With respect to the ‘danger opinion’ the Minister (and her delegates) are counselled to consider the following factors in making a determination:

1) the nature of the offence: did the offence in question involve violence, weapons, drugs; was it a ‘sexual’ offence;
2) the circumstances of the offence: its severity, what led to its commission;
3) the sentence given: as a measure of its severity;
4) recidivism;
5) humanitarian and compassionate considerations.

Similarly, adjudicators making a ‘detention release’ decision in the context of the detention review, are guided by the following factors which ‘...should be weighed when considering whether a person is likely to be a danger to the public’:

1. The seriousness of the offences:
   - their nature (offences against the person vs. offences against property):
   - the circumstances in which they were committed; and
   - the number of offences, their frequency and the pattern of criminal activity.

2. The likelihood of re-offending:
   - the person’s criminal record;
   - association with or membership in a criminal organization
   - willingness to be rehabilitated and possibility of rehabilitation; and
   - family and community support.

Under the Act, a crime that carries a maximum sentence of ten years or more is officially designated a ‘serious’ crime. It should be emphasised that under this piece of

Legally, the decision taken is not considered to be a ‘determination’ which would have certain constitutional implications with respect to due process requirements. Instead, legally speaking, Ministerial delegates are merely ‘forming an opinion’. The status of s.70(5) vis a vis judicial review will be explored in more detail a little later in this Chapter.

Cited in Falconer and Ellis 1998:20

legislation, the mere possibility of a 10 year sentence is a determinative factor: a non-citizen may meet the 'danger' criteria for expedited deportation even if they had received a fine or a suspended sentence. 'Lesser' criminals may receive a deportation order on the basis of criminality if they a) are not a Canadian citizen and b) if the offence committed falls under a Federal statute for which the maximum available sentence is 5 years or more or if a sentence of at least 6 months has been imposed. This category of 'criminals' continues to have the right to appeal their deportation to the Immigration Appeal Division (IAD) on legal (on questions of law, fact or mixed law and fact) or equitable (humanitarian) grounds.

For the moment it is important to know that while the IAD must consider each case independently and according to its own facts, general guidelines were laid out in the Ribic decision\(^4\) which continue to be cited today and which summarise the factors to be considered with respect to the Division's equitable jurisdiction (consideration of humanitarian and compassionate grounds). These are summarized as follows:

1) seriousness of the offence
2) possibility of rehabilitation
3) length of time spent in Canada and degree of 'establishment'
4) consideration of 'dislocation' effects of deportation on family
5) family and community support
6) the degree of hardship that would be caused by deportation to country of nationality. \(^5\)

While adjudicators reviewing detentions regularly form 'danger opinions' in deciding whether or not to advise the release of someone from detention, an adjudicators' decision that a person should not be released because he or she is likely to pose a danger to the public does not mean that the person will also be found to pose a danger to the public in the opinion of the Minister or her delegates. Conversely, and more

\(^4\) Ribic v. Canada (Minister of Employment and Immigration) (unreported, August 20, 1985, I.A.B., docket no.84-9623)

provocatively, a person released from detention because they are found not to represent a danger to the public may still be deported without appeal because they are found by the Minister to represent a danger to the public. While the latter scenario is rather unlikely, it nonetheless highlights not only the subjective, interpretive and discretionary nature of the ‘danger’ decision itself but also the rather odd and inconsistent effect of the division of powers between government and administrative tribunals with respect to the making of danger opinions and the consequences which they entail. As observed in the Guidelines on Detention.

[1] It should be noted that the Minister’s opinion to the effect that a person constitutes a danger to the public is not binding on an adjudicator. The latter’s decision must be based on the adjudicator’s own [discretionary] analysis and assessment of facts of the case. Therefore, it is possible that an adjudicator orders a person’s release from detention although the Minister has issued a “danger to the public” opinion.17

Writing in 1996, Weinreb and Galati observe that the Minister was applying a very wide definition of dangerousness. In their words, “Persons who have been convicted of trafficking in small quantities of crack have been found to constitute a danger where there has been no evidence of violence or weapons. Presumably the rationale is that since crack is a dangerous drug, persons who sell it are dangerous.” Section 70(5) is implicitly premised upon the dubious assumption that ‘dangerousness’ can be assessed and predicted with at least a modicum of consistency and reliability based on the consideration of certain ‘factors’; ironically, this is the very same assumption often criticized by enforcement oriented observers in relation to the release decisions of the National Parole Board.

16 Adjudicators are cabinet-appointed members of the Immigration and Refugee Board (IRB) an independent administrative tribunal. Immigration officers with the delegated discretion to form danger opinions are Ministerial representatives, civil servants of the government.

17 IRB Guidelines, 1998: 16, fn.33

18 Weinreb and Galati, 1996:57
The racist dimensions of s.70(5) have been the subject of considerable attention and contention. The ‘danger opinion’ is intimately linked, both definitionally and in its application and enforcement, with criminal justice definitions and enforcement activities. The generally accepted research finding that whites are less likely to be stopped and caught than non-whites, as documented by the Commission on Systemic Racism19 and reasserted by criminologists. 20 means that this provision, resting as it does on the policing and enforcement practices of the criminal justice system, applies and enforces the systemic racism which traverses through the CJS.

iv) Judicial Review and the Question of Discretion

The nature and status of these guidelines have been the subject of much legal attention as lawyers have attempted to draw constitutional attention to the excessive ‘vagueness’ of the legal provision and the guidelines. However, the courts position on the legal status of Departmental guidelines has been consistent, as affirmed by Judge Strayer in Williams v The Minister of Citizenship and Immigration:

I should mention briefly the guidelines issued by the department for the guidance of officers in recommending that a minister’s opinions should be issued under subsection 70(5). It was argued that the guidelines do not adequately define and limit the grounds for a finding that a person constitutes a public danger. I would first observe...that the guidelines are not law, are not binding, and they do not purport to be exhaustive. Indeed if they did purport to be exhaustive the Minister could not so fetter her discretion. I see nothing in the guidelines that is irrelevant to the proper formation of an opinion under subsection(5) (other than perhaps, humanitarian considerations to which the respondent cannot take exception) but they can in no way be seen as a definition of the totality of


the considerations of which the Minister could properly take into account.\textsuperscript{21}

While many have challenged and continue to challenge the constitutionality of s.70(5), in the \textit{Williams} case, the Federal Court of Canada upheld the overall fairness of the public dangerousness scheme. This decision nicely encapsulates, in the particular context of the danger opinion, some of the central issues raised in the Chapter Two regarding the conventional view of the relations between law and discretion in the context of exclusionary immigration law, policy and practices. In \textit{Williams}, the Federal Court reemphasized the finding in \textit{Chiarelli}\textsuperscript{22}, a case respecting exclusion on 'security' grounds, that there is no obligation on parliament to provide any kind of appeal or discretionary relief to people who have 'violated an essential condition under which they were permitted to remain in Canada'. It thus reaffirmed once again what is perhaps the most enduring of all founding tenets of Canadian immigration policy: that entering Canada is a privilege and not a right.

\textit{Williams} goes on to argue that the deportation of a permanent resident, without appeal, on the basis of a Ministerial 'danger opinion' does not represent a breach of fundamental justice. In the opinion of the court, Williams' fundamental legal rights had not been 'invaded' by the removal of his right to appeal his deportation: "...the substitution of judicial review for a right of appeal, by virtue of the Minister forming his opinion, does not strike me as a serious effect on his rights." The Federal court refused to entertain the proposition that a Ministerial danger opinion is the equivalent of 'an arbitrary order issued by a despotic official ordering the random imprisonment or exile of otherwise innocent citizens'.

In the words of Judge Strayer J.A.:

\textsuperscript{21} Judge Strayer J.A. in Reasons for Judgement \textit{Williams v Minister of Citizenship and Immigration} April 11, 1997 A-855-96(IMM-3320-95)

\textsuperscript{22} \textit{M.E.I. v. Chiarelli} [1992] 1 S.C.R.
At worst it replaces an appeal on law and facts with judicial review, substitutes the Minister's humanitarian discretion for that of the Appeal Division, and substitutes the possibility of a judicial stay of deportation for the certainty of a statutory stay.

In Williams, the failure of s.70(5) to require the Minister to provide reasons for finding that a person constitutes a danger to the public was also challenged. That courts and tribunals should provide reasons for their decisions is a generally accepted principle of fundamental justice. However in Williams, the failure of the danger opinion legislation to compel the provision of reasons for the Minister's 'opinion' is defended on the grounds that while reasons are 'desirable', they do not constitute a 'legal duty' particularly if the decision is discretionary and if there is no statutory requirement to provide reasons. Moreover, the decision notes that regardless, the issue of reasons is not really 'properly raised by this case' because the Minister is not making a 'determination' as to potential dangerousness but rather is merely 'forming an opinion'.

Nonetheless, the decision goes on to argue that even if it is properly understood as a 'decision', judicial review of discretionary, 'subjective' decision-making is limited to 'grounds such as that the decision-maker acted in bad faith, or erred in law, or acted upon the basis of irrelevant considerations.' Interestingly, this acknowledgement of the very limited grounds for the judicial review of discretionary immigration decision-making, in this instance in the form of the 'danger opinion' does not figure at all in the judge's opinion regarding the effects of a danger opinion vis a vis the individual's fundamental rights. There is in fact no absolute 'statutory' right of non-citizens to appeal a deportation order based on a danger opinion to a superior court. Nor indeed does there exist a statutory right to appeal almost any decision under the Act, with a few exceptions. There is rather, a 'very limited right to commence a judicial review' by applying for 'leave' of the Federal Court Trial Division; only about 15% of cases obtain leave. However, the

---

23 See Weinreb and Galati 1996:16

24 Ibid.,16
judge nonetheless maintains that the removal of the right of permanent residents to appeal a deportation order to the Immigration Appeal Division of the Immigration and Refugee Board, an independent administrative tribunal, through the issuance of a danger opinion, simply 'substitutes' possibility of judicial review for the certainty of appeal on facts and law.

As noted above, the judge in *Williams* also addressed the question of whether the danger opinion provision is unconstitutionally 'vague'. Interestingly but not surprisingly Strayner argues that the provision provides sufficient direction to the Minister that both she and the courts can determine whether she is exercising her discretionary authority for purposes intended by parliament. He cautions against using the constitutional doctrine of fairness "to prevent or impede State action in furtherance of valid social objectives" and adds that "...a measure of generality also sometimes allows for greater respect for fundamental rights, since circumstances that would not justify the invalidation of a more precise enactment may be accommodated through the application of a more general one." Hence, the primary defence of the 'generality' of the provision (and thus the wide scope of discretion which it entails) rests upon the importance of not impeding the 'valid social objectives' pursued by Parliament through the legislation. The guiding 'objective' is the exclusion of undesirable criminal non-citizens; 'getting tough' on 'criminal immigrants' is thus taken as an unproblematically valid social objective. Only secondarily is the possible humanitarian consequence of this vagueness mentioned and, as described above, the court had already cast some judicial doubt on the appropriateness of these 'humanitarian' considerations in the forming of a danger opinion.

v) Convention Refugees, 'Refoulement' and s.70(5): 'Deserving Victims' or 'Undesirable' Criminals'.

The question of the 'humanitarian' implications of the provision comes into stark relief when the implications of s.70(5) are considered with respect to refugees. In its consideration as to whether s.70(5) 'engages interests affecting liberty and/or security of the person', the court acknowledged the unsettled nature of the question. It suggested that
it is unclear whether the question of liberty and security of the person is only relevant in
the context of the deportation of a refugee claimant who “by definition” would be able to
“assert a potential danger to himself in returning home”. As put succinctly by the judge in
Williams.

Without purporting to decide the question in respect to refugees, I have
difficulty understanding how the refusal of a discretionary exemption from
a lawful deportation order, as applied to a non-refugee who has no legal
right to be in the country, must be seen as involving a deprivation of
liberty. Unless ‘liberty’ is taken to include the freedom to be anywhere on
wishes, regardless of the law, how can it be ‘deprived’ by the lawful
execution of a deportation order.

Canada has an obligation under the 1951 United Nations Convention Relating to
the Status of Refugees\(^25\) not to return to the country from which s/he had fled. any refugee
whose life or freedom would thereby be threatened for reasons of race, religion,
nationality, political opinion or membership in a particular social group. It is nevertheless
the case that Canadian immigration law and policy often now gives precedent to
exclusionary objectives over these humanitarian international obligations.\(^26\)

Handbook on Procedures and Criteria for Determining Refugee Status Under the 1951

\(^{26}\) For detailed discussion of the development of Canadian refugee policy see for example.
the following major works: Howard Adelman (ed.) Refugee Policy: Canada and the
United States Toronto: York Lanes Press, 1991; Howard Adelman, Allan Borowski,
Meyer Burnstein and Lois Foster (eds) Immigration and Refugee Policy: Canada and
Australia Compared (Volumes 1 and 2) Toronto: University of Toronto Press. 1994:
Toronto: York Lanes Press, 1990; Gerald E. Dirks, Canada’s Refugee Policy:
1977:180-2; Gerald E. Dirks Controversy and Complexity: Canadian Immigration Policy
During the 1980s, Montreal: 1995 (pp.60-96); Alan Nash (ed) Human Rights and
Refugees under International Law: Proceedings of a Conference held in Montreal
Inclusionary humanitarian discourses in the context of refugee determination produce and act upon the liberal figure of the ‘deserving victim’: in this case, the genuine victim of state sanctioned persecution. Exclusionary ‘criminality’ and ‘danger’ discourses produce and act upon the undeserving and/or undesirable criminal. The right to asylum in Canada of an individual found to be a Convention Refugee (a genuine and therefore ‘deserving’ victim) under the Convention is completely negated should there be a subsequent finding by a Ministerial delegate under s.70(5) that the individual is also a ‘danger to the public’ (an ‘undesirable’ criminal): the non-citizen effectively cannot be both deserving and undesirable and remain in Canada. As outlined by Douglas Lehrer, s.70(5) of the Act does indeed apply to Convention Refugees. The International prohibition against refoulement prohibits the removal of a Convention refugee from Canada “...to a country where the person’s life or freedom would be threatened for reasons of race, religion, nationality, membership in a particular social group or political opinion unless...”, as provided for in S.53 (1) (a) of the Act, the person is a member of an inadmissible class and “...the Minister is of the opinion that the person constitutes a danger to the public in Canada.” As observed by Lehrer, “the impact of a decision under 53 (1) (a) is that a person may be forcibly returned to the country of nationality, that is the country of potential persecution - in refugee law...this is called ‘refoulement’. “27

Canada, by signing the Convention committed itself to granting asylum to those found to be genuine convention refugees. It is worthy of note that Canada did not sign this Convention until 1969, a full 17 years after it was drawn up. This delay is particularly troubling given that Canada had lobbied hard for the Convention’s promulgation. The


reason Canada did not sign it until the "vast majority of countries beyond the developing world had long been signatories"\textsuperscript{28}, was the fear that signing the 	extit{Convention} would undermine Canada's ability to deport refugees on security grounds.\textsuperscript{29}

This concern seems to have been completely unfounded. As observed by Whitaker, the 	extit{Convention} "provides escape clauses for signatory states that wish to override refugee claims for reasons of national security."\textsuperscript{30} Article 32 of the Refugee 	extit{Convention} addresses the issue of 'expulsion', it reads: "1. The Contracting States shall not expel a refugee lawfully in their territory save on grounds of national security or public order." The 	extit{Convention} refugee loses their right to protection against expulsion, even to a place which presents a threat to their life or freedom, if they are deemed to be a 'danger to the public' or a 'security' threat. While the 	extit{Convention} does stipulate in Article 32.2 that this exclusionary decision must be made in accordance with 'due process of law', exceptions to this requirement are forgiven if there are 'compelling reasons of national security' which require otherwise. Article 33 of the U.N. 	extit{Convention} sets out the prohibition against refoulement:

1. No Contracting State shall expel or return ("refouler") a refugee in any manner whatsoever to the frontiers of territories where his [sic] life or freedom would be threatened on account of his [sic] race, religion, nationality or membership of a particular social group or opinion.

2. The benefit of the present provision may not, however, be claimed by a refugee of whom there are reasonable grounds for regarding as a danger to the security of the country in which he [sic] is, or who, having been convicted of a serious crime, constitutes a danger to the community of that country. (emphasis added).

Inclusionary humanitarian discourses have a far shorter history in Canada than do exclusionary discourses and in the contest between inclusionary humanitarian discourses

\textsuperscript{28} Dirks. 1977:180-2


\textsuperscript{30} Ibid..p.292
and exclusionary danger discourses. Danger discourses have tended to be dominant. The very same individual may be constituted almost simultaneously in two different governmental contexts as both a ‘deserving’ innocent victim and as an ‘undesirable’ dangerous criminal. But it is the undeserving dangerous criminal who ultimately is *acted upon* through the enforcement of the exclusionary provisions of the Act: arrest, detention and deportation.

As described in Chapter Three, inclusionary legal and humanitarian discourses relating to international human rights and liberal legal notions of natural justice and due process gained currency in the 1960s when they were officially embedded in the legislation and apparatus of on-shore refugee determination and deportation and sponsorship appeals. Inclusionary humanitarian discourses constituted the figure of the deserving genuine refugee and entailed the creation and development of a new set of governmental apparatuses to identify, verify and act upon this deserving figure. This discursive development was indeed significant and continues to entail an extensive range of political, social, economic and legal efforts, energies and expenditures relating to international and national refugee issues. However, as described in Chapter 4, over the last two decades there is evidence of a discursive shift in the governance of refugee determination: from a concern to provide protection for those who are *deserving* of asylum owing to ‘a well-founded fear of persecution’ in their country of nationality to a concern to identify and exclude those who are *undeserving* of Canada’s protection.

It is nonetheless the case that inclusionary humanitarian discourses continue to have official political, legal and social currency. Anyone facing deportation may apply for a ‘humanitarian and compassionate review’ (in 1999 the cost for making this application was $500.00) in which the Minister (through her delegates) has the case by case discretion to override all exclusionary decisions already made under the legislative provisions of *Act* and allow an individual already in Canada to stay on ‘humanitarian and
compassionate’ grounds. The Minister may also allow the ‘discretionary entry’ (30 day permit) of individuals seeking entry to Canada who would be otherwise inadmissible.31

It is still widely accepted, even by Canada’s harshest critics of its ‘liberal’ and ‘generous’ immigration and refugee systems, that Canada has an international humanitarian obligation to provide asylum to ‘genuine’ Convention refugees (although clearly most would prefer that the Canadian government select them off shore). and Canada does still grant full citizenship to Convention refugees. unlike many European countries. Similarly, liberal legal discourses relating to the critical importance of individual rights, equality, due process and the basic principles of natural justice also continue to inflect administrative decision-making in this field as is evident in the continuing existence of complicated legal and administrative provisions for Ministerial permit applications, judicial reviews and administrative appeals and in rules governing procedural matters. However, as described above, both inclusionary humanitarian and egalitarian liberal legal discourses stand down and step back when contested by exclusionary criminality and danger discourses.

vi) Conventional Critiques of s.70(5)

a) The Humanitarian Liberal Welfarist Critique: The Corruption of Discretion as a ‘Quality of Mercy’.

The passage of Bill C-44 sparked immediate outrage among community advocates and lawyers working on behalf of new immigrants and refugees. They had long been troubled by the increasing ‘criminalization’ of non-citizens. Scholarly critiques of the legislation have now also begun to emerge. Many community advocates have tended to

31 The issuance of discretionary Ministerial permits to facilitate the entry of individuals who are otherwise inadmissible under the Act for reasons of criminality is indeed analytically revealing. These temporary exclusionary permits are much more closely linked with the conventional understanding of discretion as a ‘mitigating’ quality of mercy. There have been recurring concerns expressed about the issuance of these permits, particularly when they facilitate the admission to Canada of individuals who were deemed inadmissible for reasons of criminality. These permits will be discussed further in Chapter 7.
regard the change as yet another instance of the hardening of governmental hearts regarding immigrants and refugees; in this they give expression to the welfarist/humanitarian liberal view of discretion. As shrewdly observed by eminent Canadian social commentator, Ursula Franklin,

Discretion has shifted from being a humanitarian, quality of mercy to being a license to think the worst. Discretion was initially intended to mitigate, to allow the context of an individual case to be of benefit to the recipient whereas now discretion protects the decision-maker's 'freedom to choose': it is today a quality of protecting officials from critical questions and accountability. Whereas discretion used to be founded upon an intent to 'help', it now justifies doubly harsh decisions. Discretion used to imply 'if in doubt assume the best' and now it means 'if in doubt assume the worst case scenario and act accordingly'. This is not the intent of discretion. The new Immigration Act (currently in the works in Canada) ought to specify that discretionary power ought to be used to mitigate**.

In this view, the inherent benevolence of discretion has been corrupted and discretionary processes have been hijacked from their original humanitarian objectives and has been commandeered in the name of 'getting tough'. Discretion has been conventionally regarded as a essentially inclusionary, humanitarian and desirable quality of mercy. And, as exemplified in the brief discussion of the Williams case provided above, discretion does operate to protect decision-makers from legal and political scrutiny. However, amendments such as Bill C-44 exemplify the harsh and exclusionary uses of discretion which, in the contemporary context, aim to 'crack down' and 'get tough' on 'criminal immigrants'.

To argue that the original, 'real' intent of discretion is benevolent, compassionate and inclusive, turns an analytical blind eye on the long-standing historical exclusionary uses of discretion. As demonstrated in Chapter 3, the increasing discursive dominance of humanitarian and liberal legal discourses in the post-war period did indeed provide the governmental rationale for a series of significant ('liberal' and 'progressive') changes to immigration and refugee law and policy in the 1960s and 70s. However, even in the

---

** Ursula Franklin in conversation January, 1999
context of those decades, humanitarian and liberal legal discourses never displaced the longstanding preservation of discretionary power to facilitate the exclusion of ‘undesirables’. Thus, to regard discretion as originally and inherently inclusionary, humanitarian and compassionate compels one to ignore or explain away the reality that discretion has always also been employed in the exclusionary work of Canadian immigration law and policy.

In the liberal welfarist view, discretion tends to be most commonly associated with the ‘humanitarian and compassionate review’, a provision that is, at least in its official rationale, inclusive and compassionate. As discussed in Chapter Three, while the 1960s has frequently been imagined as a watershed, a period of social and legal upheaval, a period of ‘progress’ and ‘liberalization’, even a revolution of sorts, with respect to the exclusionary uses of discretionary power, this ‘revolution’ was only partial. While it challenged the nature of exclusionary procedures and instilled an attention to ‘fairness’ and ‘equality’ in these procedures, it did not question the basic need for exclusionary discretionary authority to ‘protect against threats’, nor did it displace the primacy of this objective over any other inclusionary, humanitarian inclinations. Only when humanitarianism and compassion and liberal legality stand alone, uncontested by security or criminality or danger discourses, is it likely to prevail; and this, in and of itself, is not a new development.

b) The Liberal Legal Critique: The Threat of ‘Unfettered’ Ministerial Discretion

The liberal legal critique of s.70(5) is characterized by a central and guiding preoccupation with the effects of this legislation on the legal ideals of individual justice and notions of fairness. Liberal legal scholars by in large are not troubled by the basic

33 There is a persuasive case that the so-called ‘humanitarian and compassionate review’ is a misnomer. See for example, Anna Pratt “New Immigrant and Refugee Battered Women: The Intersection of Immigration and Criminal Justice Policy” in *Wife Assault and the Criminal Justice System* Mariana Valverde et al (eds) University of Toronto. Centre of Criminology, Toronto:1995:85-103
existence of discretionary authority to exclude 'undesirables' on the basis of criminality and danger discourses: they are primarily concerned with the degree to which the processes of exclusion respect the principles of fundamental justice and individual rights. As discussed in Chapter Two, in this view, exclusionary discretion is a problem only when it is inadequately constrained, checked and/or guided by legal or legalistic rules.

Procedurally, the Bill C-44 regime provides neither coherent discretionary structure nor adequate checking. There is no notice. There is no hearing of any kind (in fact, the point of the exercise is exactly the opposite - to remove the chance of a hearing)...At no stage is there a requirement to consult with another executive member...It is. in its lack of explicative detail and absence of any kind of external involvement in making the decision, much more an exercise in absolute discretion ... Ultimately, the process is pure discretion with no objectifying factors.34

The liberal legal paradigm redefines social, political and economic issues into legal issues: unchecked discretion is a legal problem in that it poses a threat to legality by introducing the risk of personal prejudice and political interference in decision-making and by producing 'arbitrariness'. This is offensive to and contravenes liberal legal principles, most obviously those relating to individual rights. The solution required, therefore, is the legal regulation of discretion. As put by the Canadian Bar Association (CBA): "Political determinations are unpredictable, inconsistent and risky. Accessibility to the Minister and the Minister’s delegate, for all parties involved, may become a significant and unfair factor." 35

A good and recent example of the nature of the liberal legal critique of s.70(5) is argued by the Richard Haigh and Jim Smith in their recent article.36 What distinguishes this critique from the welfarist liberal view described above is their a direct


35 CBA 1994:6

36 Haigh and Smith, 1998
acknowledgement that the changes made in the 1960s, in particular the creation of the Immigration Appeal Board in 1967 which replaced Ministerial discretionary powers as the final authority on deportations and sponsorships, were from their inception not only concerned to preserve the legal rights of individuals subject to either negative sponsorship application decisions or deportation orders, but was also and equally concerned to ‘ensure public safety’ and national security.

Haigh and Smith cite an early decision of the Board which expresses an unusually candid view of Canadian immigration policy. In *Chirwa v. Minister of Manpower and Immigration* (1969), it was held that discretion “was not intended [by Parliament] to be applied so widely as to destroy the essentially exclusionary nature of the Immigration Act and its Regulations.” (emphasis added). The authors argue that from its inception, the Board was always concerned to find ways to structure and guide their discretion in such a way as to take cognisance of the legal rights of the individual without jeopardizing the state’s interest in ensuring public safety. Ultimately, Haigh and Smith argue that the Board was better equipped to make these (necessary) exclusionary decisions in accordance with the principles of fundamental justice and fairness than the Minister.

This view was echoed by the CBA in their submission *Bill C-44*. The CBA was troubled by the fact that “[a]lmost every provision of the Bill seeks to diminish the existing rights of individuals within the purview of the *Immigration Act* or regulations.” The CBA was then, and continues to be, critical of the ‘procedural’ dimensions of the Ministerial ‘danger opinion’, noting that “[r]emoval decisions as rendered by an immigration official rather than an independent tribunal, even on the basis of Ministerial

---

37 In fairness to the authors, they do incorporate some consideration of social and political factors into their predominantly legal analyses of s.70(5).

38 Quoted in Haigh and Smith 1998:271

39 In this, their objections to the use of ‘unchecked’ Ministerial discretion to exclude rest on virtually the same (legal) grounds that were asserted by those who called for the drastic curtailment of Ministerial discretion to exclude under the 1952 *Act*, as was discussed in Chapter 3.
guidelines. may suffer from lack or procedural formality, sufficiency of evidence, inconsistency, and absence of accountability."\textsuperscript{40} However, as with the liberal legal critique discussed above, the concerns are with form, not substance. The common argument is that the IAD is better situated to make exclusionary danger decisions than the Minister (which means her delegates). The underlying exclusionary objective of the provision is not \textit{legally} troubling - indeed it is legally justified. As put in the CBA submission:

The existing structure of review mechanisms works well. Failings of the system to identify and remove appropriate residents may be best addressed through enhancement of the investigation and enforcement departments of Citizenship and Immigration and through the continued appointment of qualified members to the Appeal Division.\textsuperscript{41}

Haigh and Smith compare s.70(5) with two other legislative mechanisms that seek to govern potential "danger". Most relevant for this analysis is the "security certificate", the issuing of which can limit appeals, deny access to refugee claim determination, justify detention and facilitate and expedite deportation. The authors argue that the procedures for the finding that a person poses or is likely to pose a threat to the security of the nation are, by virtue of legal checks, less offensive than those associated with the issuing of a "danger opinion". In 1984, the Canadian Security Intelligence Service (CSIS) Act was proclaimed after considerable debate. This Act gave CSIS responsibility over national security issues, a responsibility previously held by the RCMP.\textsuperscript{42} The CSIS Act also created the Security Intelligence Review Committee (SIRC). This committee "...was to

\textsuperscript{40} CBA 1994:4

\textsuperscript{41} Ibid.,5

\textsuperscript{42} It should be noted that while names, ranks and designations were changed and while "the red serge and riding boots have gone, 95% of the men (and very few women) who woke up and started the day on July 1, 1984 as agents of the Canadian Security Intelligence Service were exactly the same people who had gone to bed the night before as members of the Royal Canadian Mounted Police Security Service." Whittaker. \textit{Double Standard} 1987:282-283
oversee the operations of CSIS and to hear individual complaints against any of its activities or its security assessments. Both the Immigration Act and the Citizenship Act were amended to provide hearings before SIRC in cases of security rejections in immigration and citizenship applications." 43 The Act provides that the Minister and the Solicitor General may make security reports to SIRC. SIRC holds a hearing in which the person in question is represented and may make a submission on their own behalf. They are not, however, provided with any comprehensive information about the nature of the evidence against them, due to "all the secrecy and governmental discretion associated with the term 'national security'" 44. While clearly the question of legal checks on Ministerial discretion to exclude is a compelling one, the focus again is on ‘procedural’ rather than ‘substantive’ justice.

c) A ‘Radical’ Critique: Ministerial Discretion, Exclusion and Racism

In contrast to the liberal welfarist critique of the corruption of humanitarian discretion evidenced by s.70(5) and in contrast to the liberal legal view which is preoccupied with the protection of individual (legal) rights and ‘procedural’ justice, a more radical critique of s.70(5) focuses on its essential racist dimensions and applications. Discretion in this view is regarded as an ‘instrument of’, or less deterministically as a ‘vehicle for’, state sanctioned racism in immigration exclusions. The contemporary preoccupation with ‘criminality’ is read as a thinly veiled, socially legitimate guise for officially delegitimized systemic and individual racism. In this view, S.70(5), and the sweeping, largely ‘unchecked’ Ministerial discretion which it entails, have opened up the possibility of the targeting of black Jamaican males for deportation:

43 Ibid., 281
44 Ibid., 282
"The unfettered discretion available to the Minister in issuing opinions opens the door for colour profiling and the targeting of certain groups." 45

In this more radical approach, the analytical danger of reductionism is ever present, despite concerted efforts to avoid its temptation. The result is that it is often unclear whether the racist effects of the legislation are regarded as the result of a racist state agenda or as unintended negative consequences.

The article by Falconer and Ellis is illustrative in this regard. Ministerial discretion sanctioned under s.70(5) is variously represented as either merely 'opening the door' to racism, as 'facilitating' racism, or finally as being racist 'in purpose and effect'. For example the authors argue that statistics show that "...the Immigration Department and its officers are operating under a system of criminal profiling that seems to target black Jamaicans in a manner that defines them more commonly as dangers to the public than any other immigrant population in Ontario, leading to their deportation in record numbers." 46 This rather ambiguous observation ('seems to target') is expressed in more definitive terms a little later on: "These statistics bear out that a government initiative has been in place since 1995 to target a specific racial group with the specific aim of cleansing the community of those perceived as a 'danger to the public.'" (emphasis added) 47 And finally, at the end of the paper, the authors refer in no uncertain terms to "...the agenda of the state in targeting members of a particular racial community." 48

The radical critique of s.70(5) should be welcomed for its attention to historical context and to substantive justice. Moreover, given the close definitional link between the Ministerial 'danger opinion' and the criminal justice determinations of criminality on which the danger opinion is largely based, and considering the now widely accepted

45 Julian Falconer and Carmen Ellis “Colour Profiling: The Ultimate Just-Desserts” Draft. 1999:21

46 Ibid.. p.23

47 Ibid.. p.24

48 Ibid.. p.26
findings that criminal justice enforcement, particularly the ‘discretionary’ components of it, gives expression to and applies systemic and individual forms of racism. The analytical focus on the racist dimensions of s.70(5) is important and compelling. This is all the more true when one takes into account the longstanding and potent historical influence of previously legitimate and explicit racist ideals, historically posed explicitly in terms of social or national ‘purity’, which contributed to and justified the development of exclusionary Canadian immigration law and policy. As described by Valverde in her study of moral reform in English Canada in the period 1885-1925.

...social purity had a clear racial and ethnic organization. The ‘whiteness’ favoured by the movement was not merely spiritual but also designated...a skin colour... ‘Racial purity’ is a phrase that appears but seldom in the texts studied, but the concept underlies common phrases such as ‘national purity’ or ‘national health’.49

The more radical view of immigration exclusions and discretionary power tends to view discretionary immigration exclusions as necessarily and inevitably racist. Explanatory reductionism is implied in the more radical critique by the lack of attention it pays to other contributing and influential factors. As a result it conveys the impression that racism is the underlying structure that ultimately governs and determines all discretionary immigration exclusions.

This view is compelling, but partial. While racism has always played a central role in the development and application of exclusionary Canadian immigration law and policy, the influence of other factors over the years must surely complicate the analytical picture. The historical development of Canadian immigration law and policy reveals that a diverse range of historically specific factors, from race to ideology to morality, to economics and labour market issues, to pragmatic concerns relating to the exigencies of electoral politics, to shifts in global migration patterns and other international conditions, have contributed to the development and enforcement of law and policy in this field.

49 Mariana Valverde The Age of Light Soap and Water: Moral Reform in English Canada, 1885-1925 Toronto: McClelland and Stewart, 1991:32
Therefore, while the current linkage between criminality, discretion and race is particularly profound in the enforcement of immigration exclusions, as is suggested by the large number of Carribean blacks who have been the subject of s.70(5). and while certainly race has always been a central factor in the development of Canadian immigration law and policy, the approach taken here focuses on the processes and practices of contemporary exclusions; exclusions which are no doubt ‘raced’, but which theoretically, are not exclusively or inevitably so. That being said, in the examination of non-legal enforcement initiatives in the governance of exclusionary immigration contained in the following chapter, the startling extent to which dominant contemporary constructions of criminality are indeed heavily ‘raced’ will be abundantly clear.

VI) Conclusion

This chapter has argued that Bill C-44 provides a contemporary example of the guiding primacy of the logic of criminality in the governance of immigration exclusions in Canada in the 1990s. S.70(5) of the 1976 Act is read as evidencing a major shift in the governing logics of immigration exclusions, from (ideological) ‘security’ to ‘criminality’. Liberal legal scholars Haigh and Smith, troubled like many by this legislative development and cognizant of the extreme and sustained opposition which parallel tracts of ‘unchecked’ exclusionary discretion has elicited in the past, sought to address the question: “Why then do we have the appearance of such legislation now, in the face of historical antipathy...?”

This is indeed a fair and interesting question. Haigh and Smith mark it down to the political and social impact of sensational media reports on the Leimonis and Baylis murders. It is suggested here that this apparent historical ‘antipathy’ extended only to the formal legal dimensions of exclusionary procedures and not to the substance of exclusions, unless they were overtly and explicitly contrary to liberal legal ideals of equality. The extreme infringement of fundamental liberal legal principles entailed by and

50 Haigh and Smith 1998:283
through the recent enactment of *Bill C-44* and s.70(5) has to do with a more general decline in the discursive dominance of governing liberal rationales and the rise of neo-liberalism over late 1970s and 1980s. The discursive dominance of liberal legality and of liberal humanitarianism has always had a fragile, tenuous and inconsistent impact on exclusionary immigration decision-making. As argued in chapters 4 and 5, this already shaky influence has further declined with the rise of punitive neo-liberal rationales, mobilized in part through central discursive preoccupations with 'fraud' and 'abuse' and 'deserving victims' and by the recent emergence of the governing logic of 'criminality' and 'public safety'. It is in this sense then that this thesis addresses the questions 'why this?' and 'why now?'

The next chapter turns to the question 'how?' In the present and previous chapter the argument has been that in the 1990s, new immigrants and refugees have been incrementally and steadily re-presented and acted upon as a frauds and criminals and as threats to public safety, to administrative systems and to national security. It is this redefinition that facilitated and legitimated the punitive legal response of Bill C-44. The following chapter takes the argument a step further, arguing that immigration policy and practice have become, to an unprecedented degree, a primary instrument of crime control. It examines the proliferation of non-legal enforcement oriented initiatives since 1994 which act upon the 'problem of criminal immigrants/refugees'.
Chapter Seven

"Criminals First": Governing Immigration Through Crime in the 1990s

1) Introduction

Throughout the 1990s, both new immigrants and refugees and exclusionary immigration law and policy have been increasingly governed by and through crime. This proposition refers to much more than merely the legal expression and enforcement of danger and criminality discourses as discussed in the previous chapter. What follows, will demonstrate the range of public and political concerns about criminality/security issues in the context of immigration have underpinned a broad range of institutional, organizational, procedural, technological and knowledge-based initiatives, networks and practices that produce and reproduce the ‘problem’ of the criminal immigrant acting upon this problem.¹ The extent to which the ‘problem’ has been acted upon legally, politically and socially and the number of agencies, agents and resources that have been enlisted in the ‘fight’ against it illustrate the degree to which the ‘problem’ of criminality

---

¹ It is worth repeating that these extensive and sustained initiatives to get tough on crime and criminality (both in the context of criminal justice and immigration) bear no reasonable relationship to any statistical evidence available pertaining to the actual size of the criminal threat posed. The Canadian Council for Refugees (CCR) points out that “There is no established connection between immigration and crime. Immigrants are actually less likely to commit major crimes than the Canadian-born and are under represented in the prison population.” Citing the findings of Derrick Thomas, the CCR notes that “[I]n 1991, 20.5% percent of the Canadian population older than 15 had been born outside the country, while only 11.9 percent of the total prison population were foreign born.” (CCR, “Facing Facts: Myths and Misconceptions about Refugees and Immigrants in Canada” mimeo. 1994:4) As observed by Matthew Yaeger, author of a comprehensive review of the empirical literature on this question, despite the fact that research has consistently found that the foreign born commit fewer crimes than the native born in Canada, “It would be felicitous to be able to record that current evidence suggests we ought to be able to dispense with the immigrant criminal myth. However, as one reviewer put it, such popular demonologies seem to be able to survive any amount of exorcism.” Matthew Yeager “Immigrants and Criminality: A Meta-Survey” Prepared for Strategic Policy, Planning and Research and Metropolis Project, Citizenship and Immigration Canada, January 1996:2
has become the central preoccupation of exclusionary immigration policy in the 1990s. Indeed, this chapter advances the argument that Immigration itself has become a central and largely taken for granted mechanism of national crime control.

The previous chapter described the 1994 extensively covered murders of a white woman and a police officer in Toronto that had triggered wide ranging and immediate enforcement-oriented initiatives. While the murders and the heightened panic which ensued gave rise to increased research activity and scholarly interest in the subjects of ‘race’, ‘crime’ and the criminal justice system, surprisingly little attention was paid to the ‘crime control’ developments of immigration law, policy and practice. It is to these developments that this discussion now turns.

II) Administrative Policy Reforms
i) ’Criminals First’ Detention and Deportation Policy (1994)

On July 7, 1994, on the heels of the murders and the public and political outcry that followed, the Minister of Immigration announced the adoption of the ‘Criminals First’ removals policy; a major re-orientation of the priorities of Canada’s detention and deportation policies. The results of this policy change was significant. In the 1993-1994 fiscal year, 15% of removals were for reasons of criminality. In 1995-1996, this figure had more than doubled, increasing to 35%. Previously, the priority of the Department had been to remove failed (‘undeserving’) refugee claimants. As observed by Richard Flageole, Auditor Principal of the 1990 Audit of Canada Immigration.

...the balance between the removal of refugee claimants and non-refugee claimants is quite different in 1996-97 than it was in 1991-92. We had more removals of claimants; now we have removals of non-claimants. That reflects the priority of the department to switch resources from removing failed refugee claimants to removing criminals.3

---


3 Citizenship and Immigration Standing Committee Hearings. February 11, 1998:10
This shift in priority is certainly consistent with the argument of this thesis that the 1980s were characterized by a guiding governmental preoccupation with fraud and system abuse (the 'bogus' refugee) and that the 1990s witnessed a shift towards criminals and criminality as the guiding governing rationale. The following table provides the numbers of criminal and non-criminal removals from Ontario and Canada from 1993 to 1997.

Table 1: Removals from Canada and Ontario, 1993-1997

<table>
<thead>
<tr>
<th></th>
<th>ONTARIO</th>
<th></th>
<th>CANADA</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Criminal*</td>
<td>Non-</td>
<td>Total***</td>
<td>Criminal</td>
<td>Non-Criminal</td>
</tr>
<tr>
<td>1997</td>
<td>940 (27%)</td>
<td>3460*</td>
<td>3460</td>
<td>1446</td>
<td>7968</td>
</tr>
<tr>
<td>1996</td>
<td>997 (29%)</td>
<td>2491</td>
<td>3488</td>
<td>1506</td>
<td>4951</td>
</tr>
<tr>
<td>1995</td>
<td>1070 (38%)</td>
<td>1756</td>
<td>2826</td>
<td>1756</td>
<td>3042</td>
</tr>
<tr>
<td>1994</td>
<td>1057 (38%)</td>
<td>1738</td>
<td>2795</td>
<td>1738</td>
<td>4700</td>
</tr>
<tr>
<td>1993</td>
<td>n/a</td>
<td>n/a</td>
<td>4826</td>
<td>1200</td>
<td>7096</td>
</tr>
</tbody>
</table>

* Criminal Removals are those stemming from criminal convictions. They may be, but are not necessarily, recipients of a Ministerial danger opinion. The bracketed figures represent the numbers of criminal removals as a percentage of the total removals.

** Non-Criminal Removals include failed refugee claimants and others who have violated the Immigration Act through, for example, overstaying their visa, working illegally or those who have been refused admission.

*** The numbers may not add up because some persons removed by have been counted in several categories. According to Immigration officials, the statistic try to capture 'primarily' the last documented reason for removal.

The impact of the murders is further evidenced by the response of the

---

'This table has been created from official statistics obtained from the Public Affairs Department of CIC through a response to information request in March, 24, 1997.'
Metropolitan Police Association to the shooting of Constable Todd Baylis by Clinton Gayle. After the Baylis shooting, the Association sponsored a lawsuit against the government of Canada on behalf of Todd Baylis and his family. In response to a government report which claimed that the failure to deport Gayle was a result of "systemic" failures, the Police Association stated their position succinctly:

Notwithstanding that one of the principal objectives of the Immigration Act is to maintain and protect the safety and good order of Canadian society, and notwithstanding that Gayle was known to be a commercial drug trafficker, heavily armed and prepared to use his firearm, immigration officials did nothing to carry out the deportation order. As a direct result of government inaction, incompetence, bureaucratic indifference and gross negligence, one of our own was murdered in cold blood. This tragedy will be repeated if we do not hold those responsible, accountable for their actions.5

After the shootings and Bill C-44 there was an increased preoccupation with risk, danger and criminality. However, this shift was neither smooth nor consistent. As was discussed in Chapter Two, the department, responding to pressures to limit financial expenditures, issued a new detention policy in the Fall of 1994 which directed immigration officers to detain only as a last resort. This both frustrated and angered Immigration officers who felt they were being given mixed and contradictory messages: first the Department reorganized its priorities around criminality and then directed that detention only be used as a last resort.

Lucienne Robillard, the Minister of Immigration, explained in 1998 that although most removals are not for reasons of criminality, the department does "...place a special emphasis on finding and removing criminals",

We will not tolerate the presence of people within Canada who threaten or prey upon the Canadian public...In 1995 we implemented Bill C-44 to speed the removal of dangerous offenders. We also created a 24 hour, seven day a week, immigration warrant response centre to provide support to our partners in the law enforcement community. These initiatives have

proved very successful, with more than 3,250 removals of criminals in the last two years, and they helped lay the groundwork for some of our recent successes.⁶

ii) CSIS Joins the Crackdown on Criminals (1996)

a) The Expansion of the CSIS Mandate

In January 1996, CSIS created a ‘Transnational Criminal Activity Unit’ as part of a “government-wide effort to combat this threat. This unit will be discussed in more detail below in the section on special departmental units. In brief, however, it draws on the Services’ operational and strategic analysis resources in order to collect intelligence related to transnational crime.”⁷ In its 1998 Public Report to Parliament, CSIS provides the justification and explanation for this unprecedented enhancement of its mandate such that it now, in its application, encompasses certain forms of criminality and merits its own unit. “Globalization”, CSIS explains, has created a world “virtually devoid of national borders.” And this, according to CSIS, provides vast opportunities for “members of highly sophisticated and organized criminal syndicates to pursue a complex web of lucrative legal and illegal activities worldwide.” Whereas terrorism and espionage have traditionally posed the primary ‘threat’ to Canada’s national security, with the demise of traditional cold war enemies and fears, in today’s’ changed global context, organized ‘transnational’ crime now poses a similarly serious threat. “...Transnational crime threatens various aspects of Canadian national security, law and order, the integrity of government programs and institutions, and the economy.”⁸ This expansion of CSIS’ mandate was effected without any amendment to CSIS’ legislated terms of reference.

CSIS now monitors ‘transnational’ organized criminal activity under its existing mandate to “...investigate foreign-influenced activities detrimental to Canadian interests, as set out

---


⁸ Ibid., 1 (emphasis added)
in Section 12 and 2(b) of the CSIS Act.\textsuperscript{9} This new focus of operations is "...in order to provide strategic advice to the government on how to deal with this immense problem."\textsuperscript{10}

In this way, certain forms of criminality have been reconceptualized and acted upon as threats to national security. Conversely, and no less significantly the dominant conception of 'national security' has been radically modified so as to encompass certain forms of criminality, thereby justifying more extreme, enforcement-oriented, widely discretionary exclusionary measures. CSIS emphasizes the 'quasi-corporate' level of organized crime encompassing "...large scale insurance fraud, the depletion of natural resources, environmental crime, migrant smuggling, bank fraud, gasoline tax fraud and corruption." However, organized criminals "...are still involved at the lower level with drug trafficking, prostitution, loan-sharking, illegal gambling and extortion."\textsuperscript{11} The range of criminality included in the new operational mandate of CSIS is thus considerably wide.\textsuperscript{12}

In contrast to the social and legal prohibition against the collection and publication of race/crime statistics, there is no official reluctance to identify the nationality and/or ethnicity of organized criminals. Indeed, organized criminals are officially and explicitly organized by nationality and ethnicity (with the curious exception of the category of "major outlaw motorcycle gangs"). According to CSIS, there are approximately 18 active transnational criminal organizations 'represented' in Canada, including: "Asian triads. Colombian cartels, Japanese yakusa, Jamaican posses. Mafia groups from the USA, Calabria and Sicily, Russian/Eastern European mafiyas. Nigerian

\textsuperscript{9} Ibid., 1
\textsuperscript{10} Ibid., 1
\textsuperscript{11} Ibid., 4
\textsuperscript{12} One cannot but speculate that this expansion and merging of the categories of criminality and security in the context of CSIS also serve to justify and reassert the importance of CSIS' continuing existence and activities in the post cold war period.
crime groups and major outlaw motorcycle gangs."  

This emergence of such a central and guiding preoccupation with organized crime as a threat to national security signals a shift in the way in which different modes of governance construct the state. The focus on ideologically based threats to national security which had been the central feature of governance in Canada throughout the Cold War years, was consistent with a conceptualization of the state as a political institution. In contrast, the contemporary focus on criminal threats to national security tend to construct the state as commerce; primarily concerned to promote the security of property in Canada and the advancement of (free-market) economic interests.  

This focus on cost, on the security of property and of free market economics, however vague, is nonetheless consistent with neo-liberal priorities including notably its primary focus on the economy and its related regulatory preoccupation with and criminalization of fraud and system abuse (the need to protect the ‘integrity’ of Canadian programs and institutions). According to CSIS, organized crime threatens Canadian society in multiple and fundamental ways. It threatens “...law and order [thereby] directly affecting people’s sense of security, trust, order and community - the very underpinnings of Canadian society”; it jeopardizes the integrity of government programs and institutions through corruption; it adds to enforcement costs; it further burdens the criminal justice system; it contributes to long-term social costs including drug dependency and a rise in violent crime; and finally it “...also poses a serious threat to the economic security of the nation” in that its basic activities could “...undermine the

\[CSIS. \text{"Public Reports"1998:2}\]

Indeed, the seriousness of the problem is often flagged by reference to the cost of transnational criminal activity. For example, in its 1998 Public Report, CSIS addresses the cost issue in a rather curious way. It begins by citing U.N. estimates of the cost of transnational criminal activity in developed states: 2% of the annual gross national product (GNP). Based on this estimate, CSIS observes that the “potential transnational crime related losses for Canada in 1995 “would have been about 14.8 billion dollars, based on a GNP of 742 billion.” (Emphasis added) There are no real figures whatsoever provided in this rationale.
workings of the free-market economy."¹⁵

This final threat to the ‘free-market’ economy is detailed as follows:

Due to their illegal activities, transnational crime groups have access to huge sums of money, which needs [sic] to be ‘washed’. This large scale money laundering has an impact on the operations of legitimate financial institutions that, in the long term, can go beyond the business sector with negative effects on the investment climate, tax revenues and consumer confidence. Moreover, these large amounts of money, combined with a willingness to use violence, enables transnational crime organizations to bribe, extort or coerce employees of financial institutions and governments.¹⁶

b) Immigration Security Screening

One of the duties of CSIS under the CSIS Act of 1984 is to provide government departments and agencies with security assessments on both government employees and contractors and on prospective immigrants and citizens in Canada. A ‘security assessment’ is defined in the CSIS Act as meaning “...an appraisal of the loyalty to Canada and, so far as it relates thereto, the reliability of an individual.”¹⁷ It is the latter responsibility that is of interest here. Security screening of new immigrants and refugees involves the cooperation of several government departments and agencies: CIC, Health Canada, Human Resources Development, the RCMP and CSIS. The Solicitor General and the Minister of Immigration make a joint report to the Security Intelligence Review Committee (SIRC) indicating that in their opinion the person in question represents a threat to national security. SIRC conducts a hearing and either confirms or rejects the conclusion of the initial report. SIRC’s report is forwarded to the Governor in Council which may direct the Minister to issue a ‘security certificate’ barring the person’s

¹⁵CSIS “Public Reports” 1998:3-4

¹⁶Ibid., pp.3-4

¹⁷Canadian Security Intelligence Service Act. 1984, c. 21. s.1
admission to Canada and/or mandating their deportation. 18

The advice provided by CSIS relates directly to the inadmissibility criteria contained in s.19(1) of the Immigration Act. According to CSIS, the amendments made to the inadmissibility provisions in 1992 by Bill C-86, particularly the insertion of explicit prohibitions against terrorists and members of criminal organizations, brought immigration’s inadmissibility provisions into closer alignment with CSIS’ definition of threats to the security of Canada, namely and primarily, terrorism and espionage. 19

In 1997-98 the Service reports that it processed 53,029 requests from CIC under the Immigration Screening Program. The average processing time was 24 days and approximately 51% of all cases were completed within this time frame. The remaining 49 percent averaged 73 days and less than 1% took longer than a year. 20

iii) Drugs and the Deportation of Refugee Claimants

As has been previously discussed, a person may be excluded from making a refugee claim under the Geneva Convention. Amendments to the Immigration Act in 1989 granted the Minister (and her delegates) the power to intervene in refugee claims if exclusion issues are raised. Under the Convention, a person may be excluded if they committed a serious non-political crime outside the country of refuge or if they are guilty of acts contrary to the purposes and principles of the United Nations. 21

18This is a vastly oversimplified description of the security review process. For the detailed provisions, see ss.39 - 40.2 of the Immigration Act.

19Canadian Security Intelligence Service, “Operational Programs: Security Screening”. CSIS Web Site. The complete threat definitions are found in s.2 (a.b.c.d) of the CSIS Act

20Ibid..

21The international and national intersection of drug law and policy with exclusionary immigration law and policy is both interesting and important, but, yes, beyond the scope of this discussion. The issue is raised, albeit indirectly, in Chapter 5’s discussion of the criminalization of the plant Khai in Canada in 1997.
Traditionally, neither of these provisions was interpreted as extending to drug offences. However, since the U.N. introduction of several Conventions and initiatives relating to international drug trafficking, the Toronto CIC Appeals office “...began to argue that drug traffickers should be denied protection as refugees.” The CRDD has accepted the argument, and on appeal the Federal Court has affirmed, that drug trafficking is contrary to the purposes and principles of the United Nations. As a result, as of 1997. “...drug traffickers are now routinely excluded from protection as refugees.” 22

III) Creation of Special Departmental Units

In addition to policy changes which elevated criminality to its primary place in the governance of immigration, the years 1994 to 1997 also witnessed the creation of several new departmental units/sections dedicated to the identification and removal of criminals.

i) Organized Crime Section of CIC (1994)

In 1994, CIC created its own 10 member Organized Crime Section. Officers in this section rely primarily upon the provision of the Act which allows immigration officials to refuse admission or remove any individual if there are ‘reasonable grounds’ to believe that they are a member of a criminal group that has engaged or is likely to continue to engage in a pattern of organized criminal activity (s. 19 (1)(c.2)). As put by Michel Gagné, Director of the Section, “We don’t have to show that the person participated in crime, just that the person is a member of an organized criminal group.” In quick recognition of the obvious implications of this provision Gagné added, “Some people believe the section offends the Charter, but we’ve never lost a case yet. ...when people knowingly associate with criminal groups the burden of proof is on them to show it would not be detrimental to Canada to allow them in.” 23 At a conference in February

22 This quote and the related information drawn from CIC Public Affairs CIC Update.
No.8. 1997:2-3

1997. Gagné reported that the mandate of the section is to: 1) understand organized crime as a global issue; 2) gather intelligence; 3) target overseas; 4) intercept at ports of entry; and 5) ensure that identified criminals are removed if found in Canada. According to Gagné, the criminal group or organization need not be tightly "organized," rather it may be "loosely structured." 

In 1998, the section boasted that it had developed a databank that contains "...the names and aliases of 7,000 suspected criminals, and those of their wives, children and parents." Lists of suspected organized criminals are shared between CIC, the RCMP, CSIS and international law enforcement officers. These are entered into the database which includes information on the person's physical description, criminal activities and links to criminal organizations. Immigration officers overseas and at ports of entry use this database for screening travellers to Canada. In its work, the Organized Crime Section of CIC cooperates most closely with the RCMP, their main "partner." 

In addition to exclusionary work within Canada and at ports of entry to Canada, the Organized Crime Unit is also active overseas. For example, in Hong Kong its officials screen potential immigrant and travel visa applicants. In this capacity they establish profiles, carry out police checks, check with RCMP headquarters and field intelligence. Since the creation of this section in 1994, Canada has denied visas to 114 out of 182,963 applicants from Hong Kong due to suspected triad membership. It has also denied visas to

---


26 "Canada Shuts the Door on Criminals" National Post, November 21, 1998

27 Ibid.,

807 out of 44,665 applicants from the former Soviet Union.³⁹

"Profiles" play a central part in the work of identifying potential criminals. And, as with those used by CSIS, the profiles developed by the Organized Crime Section of CIC accord with different national or ethnic classifications. There are Japanese syndicates, Chinese triads, Russian mafia, Jamaican posses, Colombian cartels, and then the (unstated and unspecified) 'white' categories of motorcycle gangs and what the CIC refers to as 'traditional organized crime' (previously known as the Mafia).³⁰ In the context of the CIC Organized Crime Section, profiles of members of Japanese syndicates (Yakuza) and Hong Kong triads and of members of organized crime groups from the former soviet union (vor v sakonye or thieves-in-law) have been centrally important as the above figures reflect: "Officers...target people who fit the profile of organized criminals: Japanese with missing fingers and tattoos could be yakuza; Russians who can't explain their sudden millionaire status and sport tattoos of an eight point star could be 'thieves-in-law'."³¹

ii) The CIC Modern War Crimes Unit (1996)

In 1996, the government established a centralized national unit within the CIC to monitor modern war crime cases. Regional CIC offices now have a designated coordinator to track cases and the Unit holds workshops and training sessions for CIC and the Department of Justice. As reviewed in Chapter 5, CIC immigration officers have three main tools in their efforts to exclude those who have committed war crimes, "offences against the laws of war applicable in international conflicts, for example carpet bombing" or those who have committed crimes against humanity, "...murder, extermination, enslavement, deportation, persecution, or inhumane acts or commissions committed against the civilian population or an identifiable group of persons, for example genocide.

³⁹Ibid..
³⁰Ibid..
³¹Ibid.
These three tools are: 1) the exclusion clause of the United Nations Convention Relating to the Status of Refugees (s.1Fa)) relating to the exclusion of war criminals and those who have committed crimes against humanity; 2) s.19 (1) (j) of the Immigration Act, adopted in 1987 which created a new inadmissible class: those who there are reasonable grounds to believe have committed war crimes or crimes against humanity; and 3) s.19 (1) (l) of the Immigration Act, also a new provision under Bill C-86 (1993) which deals with the exclusion of individuals who were senior members or officials of a government or regime that has been designated by the Minister as having committed gross human rights violations, war crimes or crimes against humanity.32

What is interesting about the modern (post- World War II) war crimes’ provisions is that they rely primarily upon humanitarian discourses yet they nonetheless mobilize characteristically punitive criminality/security concerns and entail coercive consequences. This suggests the continuing discursive power of humanitarian discourses. but perhaps even more significantly analytically. it demonstrates the way in which humanitarian discourses may be employed in both inclusionary and exclusionary regulatory efforts.

iii) CSIS Transnational Criminal Activity Unit (1996)

Following suit, and marking the significant shift in their orientation as mentioned earlier. CSIS created the Transnational Criminal Activity Unit in 1996. According to CSIS. this service was created “...as part of a government-wide effort to combat this threat.”34 The nature of the threat posed by organized criminals, as described by CSIS. is monumental; “...they attack the very fabric of life in a democratic, law-based society like


33Ibid.. pp.2-3. It is this last provision that was used against members of the Somali Community in Toronto in the mid 1990s as discussed in Chapter 5.

34CSIS “Public Reports” 1998:6
Canada. The work of CSIS' Transnational Criminal Activity Unit is distinguished from that of law enforcement agencies in that where the police collect ‘tactical intelligence’ ("short term and operational in nature, geared towards action in the field leading to arrests and prosecutions"). CSIS' Unit provides the government with “reliable information and strategic intelligence on the extent and nature of transnational crime in Canada.” Strategic intelligence is long-term in nature. and “...provides a comprehensive view of a threat environment. assesses the extent of the threat and points out which areas are at risk.”

It is reasonable to suggest that CSIS' new mandate issued out of a growing sense of the need for CSIS to reassert its importance in the post 1989 world order in which traditional cold war foes and fears are, for the most part, absent. Organized crime, now defined as a threat to national security, provides a new and expansive threat to be contained.

a) The Triad 'Menace' and Project Sidewinder

The following tale illustrates the degree to which factors, other than the existence of a clear and present danger, contribute to the emergence (construction) of different crime problems. In the mid-1990s in the midst of the flurry of enforcement oriented governmental initiatives related to crime and criminality, one of CSIS' first new projects related to organized crime was initiated. Project Sidewinder was a joint CSIS-RCPM inquiry into the activities of “what it felt were disturbing links between Chinese criminal gangs called triads operating in Canada and government officials in Beijing”. Indeed, in CIC, CSIS and Law enforcement reports on Organized Crime. these activities of triads are given prominent place; “Chinese triads are a main client...[they] are major. major

35Ibid., p. 4
36Ibid.
37Ibid.
38"The Trouble with CSIS" Globe and Mail, Wednesday December 15, 1999
criminals with large financial resources." The number, however, of triad members in Canada is in fact "very small": CIC admits "[W]e issue 60,000 visas a year in Hong Kong. Only a very small proportion are involved in triads." However, it is the position of some officials in this area that the absence of triad members in Canada indicates how well enforcement officers are doing their jobs of excluding them. Gagné, Director of the CIC Organized Crime Unit, even went so far as to depict Immigration and its "partners" in the RCMP, in CSIS and in the police forces as the "thin blue line" between an orderly and lawful Canadian society and the violence, corruption and economic devastation which would ensue were enforcement officers to be less vigilant. In Gagné words, "We are not being invaded by triads, because the police have displayed a vigorous response, but individual members are coming and we cannot let them multiply."  

Interestingly, the first major report about the threat posed by triads in Canada was circulated just before the introduction of Bill C-86, which contained sweeping amendments to the Immigration Act. A second report was circulated amongst immigration bureaucrats and RCMP officers while the Bill was still being debated (the Bill came into force on January 1, 1993). Among other changes, Bill C-86 included the new provision to exclude anybody "reasonably" suspected of being organized criminals.  

According to a newspaper article with the sensational headline "Triad Menace", these reports described the Triad menace as an "iceberg" and lamented that CIC was receiving requests from "...ruthless, vicious criminals."  

It is this 'threat' that provided the impetus and justification for Project Sidewinder. After producing a 12 page report on the results of their intelligence gathering, the Project Sidewinder team asked for additional resources to launch a formal probe. However, as reported in the above mentioned article, the project was dropped.  

---


40Ibid..

41Globe and Mail, "Triad Menace" October 10, 1993

42Ibid..
because senior CSIS officials found that it had produced a "rumour laced conspiracy theory" with little intelligence value.\textsuperscript{43}

IV) CIC "Partnerships": From Cops to Communities

The term 'partners' and variations thereof come up repeatedly in current immigration enforcement policy documents; the importance of enhancing and nurturing inter-agency cooperation, communication and exchange of information. Important ‘partners’ of CIC include: CSIS, the RCMP, ethnic communities, airlines, Customs, municipal, provincial and regional police forces, Interpol, foreign law enforcement authorities and governments.\textsuperscript{44} In light of the dominant policy objective of 'interdiction', the forging and strengthening connections with foreign authorities, national and international airlines and airports has become a primary preoccupation of CIC and their enforcement ‘partners’.

i) The RCMP Immigration and Passport National Enforcement Program

Historically and to the present the RCMP and CIC have worked very closely on the identification and exclusion (either through the inadmissibility provisions of the Act or through deportation) of criminals. The RCMP has an ‘Immigration and Passport National Enforcement Program’ which is primarily responsible for the investigation of violations of the Immigration Act. In April 1998, the program had 198 regular members posted across Canada. A major and continuing concern of this program is the "...proliferation of high quality fraudulent travel and identity documents". The program

\textsuperscript{43}The Trouble with CSIS" Globe and Mail, Wednesday December 15, 1999

\textsuperscript{44} Also important in the enforcement effort are CIC’s links with Social Services and Public Health departments. Social Services for the obvious reason of concerns about welfare fraud committed by new immigrants and refugees, and Public Health because of increasingly dominant fears about health risks posed by new immigrants and refugees carrying diseases, for example AIDS and tuberculosis. These links will not however be examined here.
reports that between April 1997 and April 1998 it seized 597 fraudulent travel or identity documents.\(^{45}\)

The RCMP submission to the 1997-98 CIC Standing Committee hearings outlined the 5 priorities of the program:

1) “Combatting criminal organizations involved in smuggling illegal migrants into Canada.”\(^{46}\) In 1997 the program undertook 888 smuggling related cases. MacDonald reported that triads and crime groups from central Europe have been identified by the program as being actively involved in people smuggling. The submission noted that “...due to our efforts over the past three years, it is estimated that more than one thousand inadmissible migrants have been prevented from arriving in Canada by ship.” In addition, the program has developed a database of close to 16,000 suspects and other related tactical information and shares this information with foreign and domestic law enforcement agencies. It also shares this information with Interpol “for the creation and circulation of information bulletins to all member countries.”

2) “Deterring Unscrupulous or illegal activity on the part of professional immigration facilitators.” The RCMP investigates cases of ‘malfeasance or corruption’ on the part of Government employees and of complaints about immigration lawyers and consultants.

3) “Criminal identification screening investigations of Convention refugee claimants arriving in Canada.” The criminal screening process provides CIC with information on inadmissible applicants for visitors’ visas, on potential permanent residents or on potential business investors; “The criminal screening process is designed to provide information to CIC that members or associates of organized crime groups, terrorists, persons with criminality or war criminals are attempting to enter Canada.” Over 26,000 convention refugee criminal identification screening files were opened in 1997. The RCMP also receives, records and maintains the classified fingerprints of all refugee claimants in Canada until the person becomes a citizen (or until the person reaches the age of eighty as provided under the Privacy Act) In April 1998 this program reported that their databank

\(^{45}\) Sergeant Ian Macdonald, Immigration, Federal and Foreign Policy Branch. RCMP. “Background on the RCMP Immigration and Passport National Enforcement Program” Submission to the Citizenship and Immigration Standing Committee hearings. April 1998

\(^{46}\) Ibid. This brief review of the five priorities of the RCMP Immigration and Passport National Enforcement Program is drawn from this RCMP submission to the hearings. All references and quotations contained in this section are drawn from this source.
contained the classified records of 165,000 refugee claimants.

4) “Criminal screening to identify organized crime groups and modern day war criminals.” There are two areas in which the program is active with “problematic groups”. once again, Asian and East European organized crime. The program reports that the screening process has identified 919 cases “fitting the East European organized crime profile.” The matching of the profile with the individual was successful half the time; “Of these cases, 50% were subsequently linked to criminality or organized crime groups.” With respect to modern day war crimes and crimes against humanity, the RCMP maintains a database on these specific types of crime and shares information from it with CIC and other government departments.

5) “Arrest of persons with serious criminal history who are the subject of an Immigration Act Warrant.” The RCMP responds to requests from CIC to assist in the actual arrest or removal of an individual. where “...in the view of both agencies, the individual poses a danger to the public due to a history of criminality, or who resides in a geographic location known to be hostile to law enforcement personnel.”

In their work, members of the Unit cooperate very closely with municipal and provincial police forces: “Metro, Peel, regional police and any others in the area. the O.P.P. - there’s real partnerships there.”

a) ‘Community Policing’ and Immigration Enforcement

The exclusionary work of Immigration has not been immune to the emergence of the idea of “community policing” which has taken hold in the context of law enforcement across the country. Sgt. Ian MacDonald expressed in an interview the importance of this relatively new mode of policing in no uncertain terms: “In my opinion, community policing is the only way to do immigration work...any other way is not successful. You need to get the communities on side.” Sgt. MacDonald explained that the RCMP in general. and the Immigration Branch in particular, was in a “transitional” period. “...from

---

47 Interview with Sgt. Ian MacDonald, Senior Policy Analyst, Immigration. Federal and Foreign Policy Branch, RCMP, April 23, 1998. The work of this task force is addressed in more detail in the section on the Metropolitan Toronto High Risk Unit which follows.
what they call traditional response-based policing to community-based policing.”

In the Immigration context, “community policing”, “creating partnerships with communities” “working with communities” etc., emerge in practice and indeed intent, as a way of facilitating a more ‘open’ relationship between immigration authorities and different ethnic communities in order, “...to facilitate open communication with the RCMP by ethnic communities about people who represent a threat or danger to everyone’s well-being.” Notwithstanding the possible challenge posed by different cultural perceptions of state authorities, for example those held by individuals from a totalitarian regime. Sgt MacDonald observed that “...we find that after they’ve been here a few years and they see what’s going on, they’re actually quite eager to come forward.” He added that this eagerness is due to the fact that, “...they see a challenge or threat to their lifestyle and when people from their country come in and they see they are criminals or war criminals or whatever, they see that as a threat to their own community. so they’re really keen. they definitely come forward.”

In recognition of the central importance of “getting the communities on side”. the Immigration section of the RCMP reaches out to communities. Where this is not possible due to lack of resources, as is the case in Vancouver, they have “communicated their message” through the community-based print and broadcast media of different ethnic communities. In order to effectively encourage community members to ‘inform’ on criminals and other ‘undesirables’, the RCMP has “focussed on a particular angle”. This ‘angle’ rests firmly and squarely on liberal notions of deservedness, increasingly dominant ideas about victims and victimhood as well as a certain amount of fear.

Consider the following comments of Sgt. MacDonald on the self-described ‘angle’ adopted by the RCMP vis a vis their dealings with ethnic communities:

So we’re focussing on the angle that just because you’re an immigrant or a refugee, you don’t deserve to be ripped off, and we want to know if you are. you don’t deserve to be a victim, you don’t deserve to be harassed.

---

48Ibid.
These discourses are mobilized with the *specific* intent of drawing out the *deserving* law-abiding members of different communities and enlisting them in the work of identifying and ultimately expelling the undeserving criminal from their communities and indeed from the country. In this instance ‘community policing’ is revealed as a technology of governing. A technology which quite literally *broadcasts* dominant discourses in the work of ‘hailing out’ (producing and reproducing) and enlisting self-identified law-abiding deserving new immigrants and refugees in the work of identifying (producing and reproducing) the undeserving⁴⁹. This constitutive Durkheimian strategy is quite consciously employed by governmental officials.⁵⁰ The other angle pursued by the RCMP in this process is “the safety issue”. While Sgt. MacDonald was quick to acknowledge that having visited a number of “these third world countries”, he understands why people want to leave, “...what we also have to focus on is the safety issue.” By this, Sgt MacDonald is referring to the safety of the people attempting to circumvent Canada’s immigration laws. He recounted several grisly stories of people dying in the course of attempting to come to Canada illegally and stressed the degree of control “alien smugglers” have over people once they have been transported here: “you’re basically their servant.”

The RCMP, CIC, CSIS and law enforcement agencies all employ ‘informants’ in their operations. Telling about this policing strategy in the context of exclusionary immigration is that it values, encourages, relies and acts upon community derived information about the identity of others. This stands in stark contrast to the unwillingness to accept as reliable community derived information about the identity of others when the context is inclusionary. More specifically, as discussed in the last chapter, CIC refused to accept any alternative community derived modes of verifying identity when faced with

⁴⁹Readers will recall the nature and uses of ‘community-derived’ information used against the Somali community discussed in Chapter 5.

⁵⁰Durkheim regarded crime as serving a positive function for ‘Society’ (substitute with ‘the Community’) in that it caused law-abiding members of society to stand together and against the criminal threat thereby enhancing solidarity.
the problem of, for example, undocumented Somalis applying for landed immigrant status. Affidavits from family members, employers and/or close friends were explicitly rejected by the government as an alternative to identity documents as were sworn statements from community elders and clergy. Government and law enforcement authorities are quick to accept and act upon community information when it is negative in that it implies exclusionary enforcement action, but appear completely unwilling to accept and act upon community information that is positive in that it implies an inclusionary response.

In his concluding remarks on community policing by the RCMP in the context of immigration, Sgt. MacDonald remarked that all in all he felt the approach was working. In his words, “[W]e’re going at it from that angle. I think its had a pretty positive effect. of course....” he added. “...Rome wasn’t built in a day.”


Sgt. MacDonald also discussed the generation and use of ‘target profiles’. When not engaging in community policing, the RCMP continues its ‘traditional response-based policing’: checking and monitoring flights. In their efforts to identify criminals as they disembark, the RCMP, in cooperation with CIC and Customs, develop and use risk profiles. Profiles are generated through monitoring and observing trends. As explained by, Sgt MacDonald.

For example, Air Canada flights from Singapore are starting to transport an increasing number of undocumented passengers....obviously noticed by Customs and Immigration...If this continues it becomes a trend. then authorities become interested in that flight. they can be there at the terminal to wait for the plane to disembark and then they develop individual person profiles or ‘target profiles’. So you may be looking for a male, Asian, 18-35 years old, well-dressed, new shoes, no luggage. You can hone it down that far.

While ‘target profiles’ differ depending on the ethnic and/or national origin of the criminal in question, they do not change much over time: “You may have a profile from
Asia. say Singapore. you may have another from Sri Lanka or another from Somalia or another from Russia or Israel. They are generally male, if they are female it is likely for prostitution.”

During the Citizenship and Immigration Standing Committee Hearings. Brian Grant, Director Program Management, CIC, first stated that one of the techniques used to prevent people from entering Canada is disembarkation checks. These, he said were not done on all flights, but that CIC “...targets certain flights.” Soon after he contradicted his statement and indeed those made by a Sgt. MacDonald. He denied that CIC targeted flights from high risk countries and suggested instead that CIC does disembarkation checks “at random.”51


Until 1995 the RCMP was responsible for responding across Canada on an ‘as needed’ basis to requests from CIC to carry out high risk arrests on its behalf. However, in 1995 in the wake of the 1994 murders52, CIC and the RCMP announced the creation of the ‘Greater Toronto High Risk Arrest Unit’, to be permanently located in the greater Toronto area. This is the only ‘fully dedicated’ arrest unit. Sgt. MacDonald explained that while they had considered other areas, Toronto was the only jurisdiction that “so clearly demonstrated a need for a full time unit of this nature.”

Sgt. MacDonald spoke about the influence of public and political pressure generated by the 1994 murders and their news coverage on the decision to create the Unit. He observed that when Constable Todd Baylis was shot, the media and the public had a “huge influence” on the decision. He noted that “after Baylis. suddenly we had a task force and with existing funds.”

Interestingly, the official justification for the provision of this service on a

51 Brian Grant. Acting Director General, Enforcement Branch. CIC, Citizenship and Immigration Standing Committee Hearings, February 5, 1998:31

52 See discussion of the context of Bill C-44 in Chapter 6
permanent basis in Toronto, was framed in terms of the need to protect the safety of CIC enforcement officers responsible for arresting individuals with outstanding deportation orders against them and related to that the understanding that certain areas of the city were known to be traditionally hostile to law enforcement. As recounted by Sgt. MacDonald:

After the Baylis shooting it was my understanding that Immigration had quite a few deportation orders. That they weren't able to action them all...because there were so many of them and people obviously weren't willing to be found. Baylis' killer was ordered deported one year prior to his involvement with the metro police. There was a real hue and cry and there were a lot of questions asked in the press and other contexts. as to why he was still here...Immigration reviewed their situation and one of the things they came up with was that these dangerous people pose a threat to their people who are trying to arrest them, and also that certain areas of cities were traditionally not friendly to enforcement officers. And again the risk factor skyrocketed. And I guess there was a certain reluctance on the part of immigration officers to go in there.

Sgt. MacDonald added that the fact that Immigration officers "don't carry any firearms at all" supports their safety concerns.

In light of these concerns, the RCMP put together a temporary task force at the request of CIC "to see if there was a need." After 6 months the temporary Task force found a "real demonstrated requirement" and became permanent unit. The Unit is staffed by 12 regular RCMP members and 5 CIC investigators. CIC does the initial assessment to determine whether the case is likely to entail a "high risk arrest." If they so find, they refer the case to the High Risk Arrest Unit which also reviews it before accepting it. When asked whether the Unit ever declines to assist on the basis of their review, Sgt. MacDonald responded that "...there is a really strong rapport between the arrest and immigration officers."

The criteria for assessing the risk level of an arrest are whether the crimes committed by the person to be arrested involved violence and whether the person lives in an area hostile to law enforcement. Sgt. MacDonald reluctantly agreed ("I don't know I'm just guessing") that the area surrounding Jane and Finch in Toronto, home to a large black
population is an example of "an area traditionally hostile to law enforcement".

### iii) Municipal Police Forces

CIC cooperation with municipal and provincial law enforcement agencies has also become increasingly important in Canada’s contemporary enforcement climate. The links are technological (integrated databank systems), organizational (co-locating of personnel) and practical (informal practices and procedures).

If somebody slated for deportation does not appear as required, a warrant is issued by CIC for that person’s arrest for the reason that they did not appear for removal. The warrant is entered into the Canadian Police Information System (CPIC) which can be accessed through the computer of every police officer’s vehicle across Canada. When the police stop somebody for whatever reason, if they turn out to be wanted under an Immigration Warrant, CIC is contacted, now with the assistance of the centralized Immigration Warrant Response Centre, and the person is arrested.

Constable Rob Purves, of the Morality Squad of the York Region Police, provided an example of the way in which informal understandings and communications operate in the context of policing new immigrants and refugees. As he explained, CIC enforcement officers will routinely communicate with police departments regarding possible criminal concerns relating to non-citizens in Canada, even when the concerns are not legally founded. Constable Purves described one such communication in which CIC contacted York Region police and told them that they should keep a special eye on a group of Jamaicans who had just entered the country as visitors with all their documents in order. CIC suspected that this group of Jamaican musicians was planning to perform for money at a named club during their visit. York region police responded by putting the club under surveillance. After last call, the police raided the bar and wrote a few tickets for alcohol related charges; however no immigration violations were detected. Clearly, at a time when immigration and criminal justice enforcement are both engaged in increasingly

---

53 Interview with Constable Rob Purves, Morality Squad. York Region Police Department. June 4, 1999
similar crime control work, systemic discrimination and selective policing are ongoing concerns.

With respect to organizational links, CIC works closely with all levels of police forces and the RCMP. In Toronto, the Joint Immigration RCMP Task force described above operates on a permanent basis and deals exclusively with criminal cases. In Vancouver, CIC has an officer who is seconded to the police department to assist on immigration cases. In Montreal, while there is no formal CIC representation, CIC does work closely with the police there as well.\textsuperscript{54}

In 1995, CIC set up the Immigration Warrant Response Centre in partnership with the RCMP and the Canadian Police Information Centre (CPIC) "...in order to assist our law enforcement partners."\textsuperscript{55} It enables immigration and police officers across Canada to "confirm immigration warrants anytime, day or night." Arrangements may also be made to have an immigration officer take custody of the person named on the warrant. This initiative is technology based:

The Warrant Centre manages a computer interface that links CPIC and the Field Operations Support Service (FOSS- Immigration's computer system). This interface allows Immigration personnel to access the CPIC system and obtain valuable enforcement information. The IWRC uses technology such as the photo phones which directly transmits photographs to assist in the confirmation of identity of individuals who have warrants issued against them.\textsuperscript{56}

iv) \textbf{Ongoing Collaborative Enforcement Initiatives}\textsuperscript{57}

---

\textsuperscript{54}Brian Grant. Acting Director General, Enforcement Branch, CIC. Citizenship and Immigration Standing Committee Hearings, November 27, 1997:6


\textsuperscript{56}CIC Public Affairs \textit{CIC Update}, No.8, April 1997:4

\textsuperscript{57}The following initiatives are included here, albeit very briefly, in order to illustrate the extent and variety of the crime control activity underway.
a) CIC-RCMP Task force on People Smuggling (1997)

As a means of acting upon the problem of people smuggling, the Ontario East area of Citizenship and Immigration established a task force with the RCMP in 1997. Its mandate is “...to detect and apprehend persons involved in the smuggling of persons across the border as well as those persons that have been or are being smuggled.” Task Force officers focus on obtaining sufficient evidence to lay charges under the Immigration Act.

b) Casino Task forces

Casino Intelligence Units have been set up by law enforcement agencies in Ontario in response to the ‘safety and security’ concerns posed by the opening of three casinos in Ontario. According to CIC, casinos attract a “multi-national criminal element.” Immigration Intelligence officers are assigned to each unit and where they work alongside representatives of the Ontario Provincial Police and the Customs Border Service. Their mandate is to “...monitor and target organized criminal groups to prevent their profiting by the casino industry, both nationally and internationally.” According to CIC, the three casinos conduct 600 investigations per year “...on various criminals and their associated organizations.” These cases involve both residents and non-residents of Canada. A permanent resident may be subject to removal if a conviction results from the investigation.

c) CSIS and CIC Airport Collaboration Project

In 1997, a pilot project was initiated at Pearson International Airport in which CSIS officers and CIC officers conduct joint interviews “...of those persons posing a potential threat to the safety and security of the country.” This initiative is attributed to...

58Public Affairs, CIC Update, April 1997

59Ibid..2

60Ibid..2
the two departments' mutual recognition of the importance of "...interdiction activities in refusing entry to those persons who pose a threat to the safety and security of Canada." \(^6\)

d) Ski-Doo and Boat Patrols

In 1997, some CIC port of entry staff who work and land border points in Ontario were part of a collaborative Customs and Ontario Provincial Police (OPP) program in which they carried out border patrols on ski-dos and boats. Accordingly, they received 50 hours of training in a formalized OPP boating course. \(^2\)

V) Airlines, Airports and Interdiction

In light of the dominant policy objective of "interdiction", the forging and strengthening of connections between CIC and foreign authorities, national and international airlines and airports has become a primary preoccupation of CIC and its law enforcement "partners". According to Sgt. MacDonald, the inability to effectively monitor all planes at all times means that CIC and the RCMP must be "multi-dimensional." So, while the Immigration Branch has "dedicated officers" at centres at the Toronto and Vancouver airports, they must also pursue other strategies. \(^3\)

One such strategy is interdiction. As has been discussed, since the 1980s, there has been an increasing preoccupation with interdiction and deterrence as policy justifications and objectives in the field of immigration exclusions. The RCMP Immigration Branch is centrally and actively involved in interdiction initiatives overseas.

\(^6\)Ibid..6

\(^2\)Ibid..6

\(^3\)

Airline liability assists the RCMP and CIC in enforcing interdiction oriented policies. Airlines are a "stakeholder" in this problem, because if they do bring in undocumented people they are subject to stiff fines. As pointed out by Sgt. MacDonald, they can also be held accountable for that person's expenses while in Canada as well as for the cost of returning them: "...so the airlines do have a real stake in keeping these people out." (Sgt. Ian MacDonald. RCMP, Interview April 23, 1998)
As explained by Sgt MacDonald,

With 125 million global migrants, the world is becoming a smaller place. That’s understood, but we find that more and more we have to work with foreign authorities to address these issues, instead of just waiting for the thing to happen in Canada when it’s just too late.64

CIC agrees: “[F]or us, a tremendous emphasis is placed on trying to screen out inadmissible persons before they get here, because we know that once they get here, they’ve got full access to our courts. and they can delay proceeding for a long period of time.”65 Lucienne Robillard, Minister of Citizenship and Immigration, stated in 1998 that interdiction efforts are of paramount importance: “The most important thing is to prevent these people from coming to Canada. That means trying to stop them before they come...part of our strategy is to work overseas in the different airports...we also work with other countries.”66

In the interdiction effort pursued by CIC, enhancing links with national and international airlines is key. In the mid-1990s, CIC introduced a new approach with airlines that resulted in greater cooperation in the policing of travellers. Fraudulent and/or missing travel and identity documents are the primary focus of many of the initiatives taken. CIC works very closely with airlines at major transit points around the world training airline employees and other authorities on how to detect fraudulent documents. for example, in Frankfurt, London, Paris, Hong Kong, Tokyo and Singapore.67 In addition to training initiatives, CIC also posts officers at international airports to actually assist foreign authorities in the detection of fake documents. Sgt. MacDonald explained

64Interview with Sgt. Ian MacDonald, April 23, 1998
66Lucienne Robillard, Minister, CIC, Citizenship and Immigration Standing Committee Hearings. November 18, 1997:29
that sending Canadian officers abroad to do this work is necessary because "...we can't expect them to look after our interests, and they don't know what they are looking for." CIC also 'insisted' that international airports check documents before getting on the plane, as opposed to relying on the disembarkation check in Canada. The CIC officers posted abroad have also been given the legal authority to hold the documents of people getting on the plane.68

The largely unacknowledged 'unintended' consequence of 'successful' interdiction efforts which aim to prevent people from getting on a plane in the first place, or from being admitted to Canada at an airport or a border port of entry, is that other avenues of entering Canada are likely to be explored. As was learned from the prohibition of alcohol in the 1920s, successful interdiction measures are likely to redirect the undesired activity in more covert, risky and dangerous directions rather than eradicate the activity. This restrictive enforcement oriented immigration policy and practices that focus primarily upon interdiction and deterrence oriented measures actually produce and reproduce some of the very problems which it is designed to address and ameliorate, including most obviously people smuggling, and other methods of system abuse. This likelihood is considered in the following observation made by Brian Grant, Director of Program Management, CIC:

...while [interdiction] is necessary in order to protect the system from being overrun, and I believe it would be...nonetheless, we are offering resistance. If we could prevent everyone from stepping on an airplane, I suspect you would have boats arriving on the coast. People will try to get here no matter which way.70

Official measurements of how successfully interdiction measures are working

68 Interview with Sgt. Ian MacDonald, April 23, 1998

69 Brian Grant. Acting Director General, Enforcement Branch. Citizenship and Immigration Standing Committee Hearings on Immigration, March 18, 1998

70 Brian Grant. Acting Director General, Enforcement Branch CIC. Citizenship and Immigration Standing Committee Hearings, February 18, 1998:27
include the number of overseas ‘interceptions’ and the total number of people arriving at Canadian airports. On both counts, tough interdiction measures do appear to be ‘working’: in 1991 the interception rate was 30% and in 1998 it was 54%. The number of people arriving at airports also dropped by half. However, as noted by Grant, the number of people arriving at our land border has increased: "...you just keep chasing after the problem." At the time of writing (January 2000) a recent series of arrivals of Chinese nationals on the west coast of Canada, smuggled across the ocean in a shipping containers, several not surviving the journey, seems likely to be a very sad example of the ‘unintended consequences’ of get tough exclusionary immigration policy.

Further concerns about intensive interdiction measures question the consequences of preventing, and thereby endangering, genuine refugees from making a refugee claim in Canada. And, despite international pronouncements on the need to ‘share the burden’ of massive international migration, tough interdiction and exclusion measures clearly indicate a desire on the part of the Canadian government to keep to an absolute minimum the numbers of all refugee claimants, ‘genuine’ or not, who are able to reach Canada and avail themselves of the on-shore refugee determination system.

Notwithstanding these concerns, CIC is clear that one of their major priorities continues to be to increase cooperation with respect to overseas interdiction. CIC has control officers in different airports around the world is convinced that these links (‘partnerships’) with foreign airlines and authorities need to be strengthened in order that Canada “doesn’t become a safe haven.”

---


72 Ibid. 8

73 As has been discussed in Chapter 4, the ‘Safe Third Country’ provision provides further strong evidence that Canada has prefers to ‘shift’ rather than ‘share’ the so-called burden of the world’s migrating populations.

74 Minister Lucienne Robillard, Citizenship and Immigration Canada. Citizenship and Immigration Standing Committee Hearings, April 29, 98
Initiatives in this regard are also technological. CIC is trying out what is called an "Advance Passenger Information System". This involves drawing information from the tickets of passengers flying to Canada and sending that information ahead. CIC officials in Canada then run the information through the available 'criminality databanks', to see "...whether we get a hit either in terms of criminality or a security concern...so we can know who's getting off the plane."75 There have also been concerted efforts to develop alternative measures that would harness and effectively employ the potential of technology. Document scanning technologies for example; "...as the technology develops it gets easier to do this sort of thing."76

Disembarkation checks by CIC officials are also being employed more frequently. CIC officers board international flights upon their arrival and prior to the disembarkation of the passengers in order to check documents. At Pearson International Airport in Toronto, it was reported that between September 1996 to January 1997, 660 disembarkation checks were completed compared with 237 checks done over the same period the previous year.77

VI) Canada Customs and Crime Control

January 26th, has been designated "International Customs Day." This tribute is designed to "recognize the contribution of the men and women of Canada Customs to the safety of Canada's communities."78 There are about 4,000 customs officers at border points and ports of entry across the country. Their responsibilities range from "collecting

---

75 Brian Grant. Citizenship and Immigration Standing Committee Hearings. February 18, 1998:20

76 Ibid. p.35

77 Public Affairs. CIC Update. April 1997:5

duties, to enforcing health regulations to investigating drug smuggling.” 79


Customs officers are represented as the “..first line of defence against drugs, contraband, and illegal firearms.” 80 In 1997, Minister of National Revenue, Jane Stewart, emphasized their crime control function: “Working with the RCMP and other domestic and international law enforcement agencies, the men and women of Revenue Canada’s customs’s’s operations have contributed greatly to keeping our communities and our streets safe.” 81 Customs officers are also responsible for carrying out the ‘primary inspection’ of travellers for CIC as discussed in Chapter Two.

In March, 1997, Minister Stewart announced the tabling of legislation (Bill C-18) which sought to expand the powers of customs officers under the Criminal Code and in May 1997 the Bill was given Royal Assent. Bill C-18 expresses the view that customs officers provide the “first response” to criminals seeking entry to Canada. As such, it vested them with expanded powers “…to arrest and detain individuals suspected of having committed offences under the Criminal Code.” When fully implemented, 3,000-3,500 of the approximately 4,000 customs officers will be “…designated to provide this first response capability.” 82

This legislation’s development and adoption was explicitly attributed to the contributions of a variety of agents and agencies including law enforcement, Canada Customs employees and most notably non-governmental lobby groups including “Canadians Against Violence Everywhere Advocating Its Termination” (CAVEAT) and

79 Ibid.

80 Ibid.

81 Ibid.

Revenue Canada explains that it has always had to balance its responsibilities of "facilitating trade, travel and tourism" with that of "maintaining a strong and credible enforcement role." It has implemented a "Smart Border" strategy. This strategy is "based on effective risk management, which allows for the speedy clearance of low-risk people and goods while at the same time keeping out undesirable merchandise and people." The expansion of the coercive powers of customs officers designated as 'first responders' is part of the latter exclusionary objective.

ii) CANPASS and the 'Low Risk' Traveller (1995)

In 1995, the Canadian government launched a "smart card" to speed customs clearance for "low risk" travellers as part of its CANPASS Program: the inclusionary dimension of the 'Smart Border' Strategy. This initiative issued out of a 1995 Accord with the United States (the "Canada-U.S. Accord on our Shared Border" February 1995). Prime Minister Chretien expressed the rationale for the initiative as follows: "we have a long history of cooperation with the United States. Revenue Canada is building on that tradition to improve the way we do business...Better service at the border will boost tourism and trade and that means more jobs in Canada." The economic rationale for inclusionary measures is of course historically long-standing and consistent. As has been argued, in the 1990s inclusionary policy objectives are opposed primarily by criminality

---

83 Ibid.
85 Ibid.
86 Ibid.
88 Ibid.
based exclusionary rationales. According to National Revenue Minister David Anderson. “Our goal is a hassle-free border for honest travellers and businesses, and a brick wall for those who try to smuggle illegal weapons, drugs or break other laws at the border.”

To be eligible for a CANPASS ‘smart card’, one must qualify as a ‘low risk’ traveller. A low risk traveller is: a citizen or permanent resident of Canada; a citizen or resident of the U.S. who meets the Canadian visitor requirements (is of “good health, no criminal or narcotic record and the ability to financially support yourself while in Canada); a citizen or resident of the US entering Canada to work or study “who meets all immigration requirements, which may include possession of a valid written authorization from an immigration officer. Conversely, you are deemed a ‘high risk’ traveller and are therefore ineligible for the CANPASS program if you: do not meet the above requirements; have a criminal record for which a pardon has not been granted; had a customs seizure within the past 5 years; have been subject to an enforcement action under customs or immigration legislation; or are inadmissible under the Immigration Act.

VII) Technological Developments

The importance of technology in the crime control work of immigration is no doubt by now abundantly clear. As has been described, it is a central part of CIC’s ongoing efforts to enhance ‘partnerships’, forge new links and share information. It also holds out the promise for many that enforcement authorities will, through the development and implementation of new technologies, be able to better identify, track, detain and exclude those who are found to pose a threat to public safety and security.

Currently CIC maintains the ‘FOSS’ system which records immigration transactions. CIC has undertaken the development of integrated processing systems, to replace the various tracking systems which had been developed in different regions and which are for different stages in the process and which are not integrated. The National

89Ibid..

90Ibid..
Case Management System would standardize client identification codes and would provide all enforcement officers across the country with access to an ‘integrated client file’ which would enable CIC to track cases throughout the process without having to jump from one system to another. This system would also include a ‘data warehouse’ which would facilitate decision-making, reporting and statistical calculations and reports. CIC is also working to enhance its links with other ‘partners’ and ‘stakeholders’. It would like to see paper interactions replaced by electronic ones and is working to “…build suitable interfaces between the various systems.” CIC would also like to see the United States and Canada integrate their tracking and information technologies so that illegal immigrants could be tracked in the U.S. and in Canada. To that end, Minister Robillard explained that, in 1997, CIC was setting up a working group with the U.S. to explore the possibilities.

CIC is also developing a CIC Explorer, an Intranet, to assist officers in Canada and internationally to access to a wide range of information which they need to draw from in their work. CIC has begun using video-conferencing for detention reviews, and is exploring the use of electronic bracelets as an alternative to detention. As already mentioned, document scanning technologies are also being explored and developed. While not new technologies, the fingerprint and the photograph are critical devices in the identification and tracking work of CIC. As put by one official, “people may have fabricated an ‘identity’, but they do have a fingerprint and a photograph. CIC has what’s referred to as a ‘positive identifier’.  

---


92 Ibid., p.7

93 Minister Lucienne Robillard, CIC, Citizenship and Immigration Standing Committee Hearings. November 18, 1997:7

94 Brian Grant, CIC, Citizenship and Immigration Standing Committee Hearings. February 26, 1998:18
As summed up by Foster:

There has simultaneously been a preoccupation with the advanced technology and science of "admission and control" procedures. Infra-red surveillance devices, magnetic strip documentation, telephonic fingerprinting, computer databanks, x-ray security at airports, electronic filing systems and automated mail in systems all serve, in a rather emphatic attempt, to regulate the flow of economic migrants and refugees from Africa, Asia, the Caribbean and Latin America.  

VIII) The Production of Knowledge on the ‘Problem’ of the Criminal Immigrant

The discursive construction and reproduction of the problem of the criminal immigrant/refugee also takes place in contexts devoted to the production and dissemination of knowledge about the ‘problem.’ This takes place in, for example, political and social forums, in committees, in public hearings and in the media.

The 1990s have witnessed a proliferation of research-oriented initiatives on different issues relating to Canada, immigration and immigrants. The relationship between immigrants, immigration and crime is one such issue that has received much attention. One illustrative example of an important, government-initiated and funded site of knowledge production on the ‘problem’ of criminal immigrants/refugees is CIC’s ‘Metropolis Project’.

i) The CIC Metropolis Project

Meyer Burstein, the Executive Head of the Project in 1997, explains that it has become apparent to governments that they need to take leadership, "...to build new institutions, to engage a broader spectrum of stakeholders and to create a shared strategic focus." In order to effectively manage both “the flow of immigrants” and the

---


“consequences of immigration for the host society”, information is required “that we do not presently have.” Governments therefore have realized that they “need to invest more heavily in knowledge.” (emphasis in original) The objectives of the Metropolis Project are to:

... improve public policy by situating knowledge at the core of decision-making. Our goals are to build a unique network of researchers and decision-makers; to create a policy thrust by continually involving governments and stakeholders in project design and problem definition; and to energize the network through conferences, policy forums and extensive informational exchanges.

Related to this endeavour, four ‘Centres of Excellence’ have been set up by CIC in Montreal, Vancouver, Toronto and Edmonton. These centres conduct research related to settlement and integration issues, with consideration extended to both the so-called ‘host’ communities and to the immigrants themselves. Burstein explains that these research parameters are so wide that the ‘stakeholders’, consisting of both “knowledge producers” and “knowledge consumers”, need to identify clear priorities to guide this research. This was the stated aim of the Justice/Immigration Domain Seminar, held by Metropolis in February, 1997.

In spite of its title, the issue of crime and immigration and criminality and immigrants provided the policy focus for this two day seminar. The issue of ‘justice’ was not much addressed; in the mix of the issues addressed during the seminar, it was ‘enforcement’ that was the dominant preoccupation. It provides a good example of a specific site in which knowledge about the ‘problem of the criminal immigrant’ and related preoccupations were produced and reproduced. It was co-sponsored by the RCMP, the Correctional Service of Canada, the Solicitor General of Canada Secretariat and the Department of Justice.

Government conference organizers in fact took great pains to acknowledge the

---

97 Ibid.
98 Ibid.
existence of, and rhetorically to refuse to reproduce dominant myths and misconceptions which inform the 'public's perception of immigrant criminality'. They also acknowledged the limits which 'polemic and partisan' debate on this issue place on policy discussions, development and decision-making. The difficulty, according to Burstein, was that few researchers have avoided the tendency to approach the issue from either the 'immigrants as victims of crime' or the 'immigrants as perpetrators' of crime. The promise of a new approach was quickly dashed when it became clear that the dichotomy was not to be unsettled by this conference, but rather would be more comprehensively reproduced: both sides of it were to be addressed, not just one. As put by Burstein, "This seminar will address both aspects of the issue...This comprehensive approach [is] reflected in the two-fold structure of the seminar." The conference was accordingly divided into two sessions: 'immigrants as participants in crime' and 'immigrants as victims of crime'. In the first session, seminar participants heard from senior representatives of KPMG Security Inc. (a private security company), the Criminal Information Service of Canada (CISC, a national police organization), the Organized Crime Unit of CIC, a sociologist on the subject of gangs and a representative from the Dutch Ministry of Justice. In the second section, three academics made presentations on hate crimes, racism in criminal justice and legal pluralism. The seminar's working groups were similarly structured by the victim/offender opposition. All workshops were open to all participants except one which was 'open to law enforcement personnel only.' The result was that very few enforcement officials worked with anybody other than other enforcement officials. Although there was a stated concern to avoid a we/they approach and to sidestep the limits of conventional oppositions and misconceptions, the very organization and structure of the seminar did little to unsettle conventional understandings and assumptions.

Another aspect of this seminar was the guiding and pervasive tendency to conflate and treat as one and the same immigration enforcement and criminal justice enforcement.

---

"Ibid."
This was identified in the executive summary of the seminar as one of the 'critical issues' identified in the 'enforcement' working group: "There has been a general failure to distinguish between immigration enforcement and criminal justice enforcement. While there may be overlapping areas of responsibility, the two systems need to be distinguished from each other."\textsuperscript{100} However, as has been argued in this thesis, there is a very good reason for the constant slippage between the two varieties of enforcement. Immigration law and policy has over the past decade been reconstituted as a primary crime control instrument. In the present analysis, there is little need to specify what agency is doing the enforcing as they are today both regarded as being in the same business of controlling crime and criminals in the interest of public order and safety.

IX) 'Good Guys' and 'Bad Guys': The Entrenchment of an Enforcement Mentality Amongst Frontline Immigration Officers

The incremental and systematic reconceptualization of immigration law and policy as a critical and central instrument of crime control is a process which took hold with the 1976 Immigration Act and has continued to the present day. The sustained preoccupation with exclusionary enforcement which this reconceptualization has both relied upon and reproduced has contributed to the entrenchment of an 'enforcement mentality' amongst front-line immigration officers. This is neither a new nor an uncommon concern. Indeed it is one that was raised frequently by representatives of non-governmental organizations during the course of the 1997-98 CIC Standing Committee Hearings. They were clearly dismayed that the recommendations made in the 1996 Tasse Report regarding this problem had not been taken seriously. In 1996, the enforcement mentality of the Enforcement Branch of CIC was identified as a serious problem which results in numerous infringements of people's rights, inappropriate uses of coercion and intimidation, inhumane practices that are illegal, borderline legal, or merely underhanded. As noted in the Tasse report, "Removal officers often act as if they believed that

\textsuperscript{100}CIC Metropolis Project “Justice/Immigration Domain Seminar” Report 1997
all individuals under a removal order were dangerous criminals, liars and dishonest. This leads to bad attitudes and the improper treatment of people.” 101 In response to the finding that there was an understanding amongst enforcement officers that they should do everything necessary to carry out a removal, and that there was little consideration given to ethics, the Tasse Report recommended a complete overhaul of CIC Enforcement’s code of ethics.102

Immigration ‘enforcement’ consists of five main components: investigation, detention, inquiry, appeals and removals. There are over 550 enforcement officers across Canada who deal with immigration ‘violations and violators.’ According to Gerry Campbell, Assistant Deputy Minister, Operations, CIC, the Enforcement Branch of CIC “...is probably one of the most difficult and complex spheres of all CIC’s activities. High profile cases often involve deportations, security certificates, and legal challenges of refusals.”103

Former frontline immigration officer, Lorne Foster, observes that “[t]here is no doubt that being on the frontline in the immigration business affects one’s view of the human panorama.”104 Foster’s description of the enforcement mentality shared by frontline officers nicely exemplifies this view:

To seasoned immigration officers the matter is clear: you either do the job of enforcement and control or you get out. There is no alternative or in between. For the frontliner, the big tribe always takes precedence over the little tribes: ...There are only “good guys” and “bad guys.” You do your damnedest to get the good guys in and get the bad guys out....This good-guy-bad-guy equation on the frontline...guides immigration


102 Ibid.

103 Gerry Campbell, Assistant Deputy Minister, Operations. CIC, Citizenship and Immigration Standing Committee Hearings, October 21, 1997:16

104 Foster Turnstile Immigration,1998:24
officer's conduct beyond explicit rules."

This enforcement mentality results in many questionable practices. Indeed, the Tasse Report itself was commissioned in October 1995 by the Minister of Immigration in response to one such practice that had been made public a month prior. A former police officer and manager of CIC in Winnipeg was accused of falsifying deportation papers to expedite the removal process. While the CIC advised the RCMP that the story “did not merit a formal investigation”, the RCMP went ahead with its own internal investigation to find out why the case had not been acted upon. They also initiated a criminal investigation. It was after this incident that the Minister appointed Tasse to review removal policies, observing that: “[I]t’s not an easy job but...forgery is unacceptable and should never be condoned.” Subsequently, two other immigration officers from Mississauga, one of whom was also a manager, were charged with forging a document also to expedite removal.

No charges were laid against the Winnipeg Immigration manager and in March 1996, the criminal charges against the two Mississauga immigration officers were dropped. While initially the Crown had wanted a conditional discharge and community service, one of the officers charged had been promoted while the case was being handled. Because her posting was in Tokyo and she could not therefore be monitored, the crown considerately agreed to drop the criminal charges and instead she and her colleague were charged under the Immigration Act with ‘making a false statement.’ According CIC, the officers were also ‘disciplined’ by the department; in fact they received a several day

---

105Ibid., 30

106CBC. ‘The National’. Transcript from September 6, 1995

107Globe and Mail. “RCMP Probe Forgery. Deportation Staff to be Questioned”. October 10, 1995

108Ibid..
suspension.109

Immigration lawyer David Matas describes the enforcement mentality as an organizational and systemic problem that issues in part from the dual purposes of the Immigration Act. inclusion and exclusion, admissions and enforcement:

If you look at the Act [it] serves two purposes. It is to allow some people in and to keep other people out...the department is bicephalous. But the individual people, they only perform one function. There’s an admissions side and an enforcement side. The enforcement people develop a group ethic and they tend to support one another, they tend to agree with what they are doing. they attend to trade practices.110

This process of sub-culture formation entails a shared perspective that informs the ethic and the practices. In the case of the ‘enforcement side’. this content resembles those now fabled days of the wild west; there are ‘good guys’ and ‘bad guys’. cops and criminals. borders and badges, cuffs and shackles, cells and exercise yards:

...the custodial ‘feeling’ or ‘desire’ of the frontline immigration officer is to wrestle the bad guy to the ground right at the nation’s turnstiles. take him for a spin and send him home...Despite all the fine print and officialese. seasoned frontline immigration officers are really only concerned with one thing - rooting out the bad guys; everything else is bureaucratic minutiae and ‘small potatoes’ and not really worth worrying about.111

This view is not a secret; as observed by Matas: “The problem is that...policy is administered by people who see themselves as protecting Canada’s borders. They see everyone in front of them as trying to trick and cheat their way into Canada. They’re very hard. and in some cases much too hard.”112


111Lorne Foster Turnstile Immigration, 1998:30

Arguably, for all intents and purposes, the perspectives, objectives, powers and practices of immigration enforcement officers today bear a striking resemblance to those shared by police officers; a resemblance which most certainly underpins the tendency to conflate the two varieties of enforcement. There are, of course, distinctions. Two are relevant here. First, immigration officers carry no firearms. However, as was reported when the CIC/RCMP Task force was first announced in July 1994. "[A]s part of the beefed up effort to catch foreign criminals, especially violent ones, many enforcement officers will be equipped with pepper spray, collapsible batons and police-style training." These innovations when coupled with the already extensive powers of immigration officers (arrest, search and seizure and detention) surely make the difference owed to the absence of guns less compelling. Second, the enforcement powers and practices of immigration officers as governed by the Immigration Act are not subject to nearly the same degree of legal, organizational and public scrutiny as are those of police officers. This reality, as discussed in Chapter Two, has lead some critics to call for an external civilian review mechanism of enforcement operations and other measures to increase accountability.

X) Conclusion

The construction of a community founded upon the expulsion of textual outlaws...requires the establishment of borders and boundaries. Beyond the boundary can be identified the outlaws, within the boundary exist the members of the community, attempting to 'lead their lives free from fear and in relative security'...The localism of crime necessitates the nomadism of law's response. Locating crime, pinning it down, becomes a major preoccupation...Where crime is located, the border can be strengthened. 114

The degree to which criminality is a rhetorically powerful 'problem' in the governance of immigration is also evidenced by the existence and use of an important


114 Young Imagining Crime 1996:12
inclzrsionary discretionary Ministerial power: the discretionary temporary entry permit. The issuance of this (non-extendable) temporary permit under s.19(3) of the Act overrules any and all exclusionary decisions previously made allows a legally inadmissible person to be enter Canada for a period not exceeding 30 days. In this, it resembles a ‘merciful’ application of discretion providing ‘relief’ from the hard edge of the law. However, it is relief with a twist. Whereas inclusionary Ministerial discretionary power is most commonly associated with the granting of ‘humanitarian and compassionate’ relief, this measure does not have anything explicitly to do with humanitarianism. Under s.19(3) of the Act, the discretion to issue a temporary entry permit is delegated to senior immigration officers and adjudicators. They are to consider three factors: 1) that the applicant is a member of an inadmissible class under s.19(2) of the Act; 2) that the purpose of the entry justifies admission; and 3) if entry is to be granted, appropriate terms and conditions are imposed for entry and for the duration of the stay in Canada.

What is of particular analytical importance with respect to this provision is that it is used to grant entry to criminals, even so called ‘serious’ ones. Clearly, the government’s prohibition against criminals is not a zero tolerance position, despite the rhetoric. The 1997 Ministerial report to the House of Commons on the issuance of discretionary entry permits, reported that over 4,000 permits had been issued in that year. Of those, 1,497 were issued to people who were otherwise inadmissible for reasons of criminality. Of those 1,497, 395 of the crimes in question were described as ‘serious’.

In the 1997-98 Citizenship and Immigration Standing Committee Hearings, this practice was the cause of considerable contention. The Reform Party of Canada sought to have a motion passed requesting the names, crimes and countries of origin of these 1,497 people. This elicited much opposition from those who were concerned about the racist uses of such information as well as about privacy issues. For the Reform party, the issuance of discretionary Ministerial permits to criminals brings the integrity of the system into question and casts doubt upon the government’s commitment to cracking

down on criminals.

For this author, however, the issuance of these permits indicates the degree to which the problem of criminality, along with the myriad of interventionist techniques which it justifies, is discursively produced. The official and political representation of immigrant and refugee crime and criminality is for the most part, unwavering in it condemnation. The Ministerial granting of relief to criminals is seemingly inconsistent with that condemnation. It has been seen time and time again that humanitarian notions of mercy or relief are consistently ‘trumped’ by criminality/danger discourses. This practice indicates a crack in the constructed armour. A quiet, backdoor provision which concedes that ‘criminals’ are not always undesirable, and that strict hard-lined prohibitions against criminal immigrants, while publicly important, are in practice are not as definitive as the rhetoric implies.

On yet another analytical level, the issuance of these discretionary permits is consistent with the form of sovereign power employed against undesirable and undeserving non-citizens in the context of contemporary immigration penalty. The sovereign’s power to be merciful necessarily accompanies the sovereign power to exact bodily punishment. Like the sovereign of the middle ages, the Minister of Immigration (and her delegated officials) are present,

...not only as the power exacting the vengeance of the law, but as the power that could suspend both law and vengeance. [the sovereign] alone must remain master. he alone could wash away the offences committee on his person; although it is true that he delegated to the courts the task of exercising his power to dispense justice, he had not transferred it: he retained it in its entirety and he could suspend the sentence or increase it at will.116

The 1990s have witnessed an unprecedented preoccupation with crime and criminals in the field of immigration. In cooperation with a vast array of ‘partners’, CIC has pursued a systematic and comprehensive agenda to get tough and crack down on

116Michel Foucault. Discipline and Punish: The Birth of the Prison. New York: Vintage Books. 1979:53. This understanding of immigration penalty as contemporary sovereign power is elaborated and exemplified in the following chapter.
criminal immigrants/refugees. The ‘criminal immigrant/refugee’, as constructed over the 1990s, threatens the security of Canada, the efficiency of the free-market economy, the integrity of state systems, and the health and safety of Canadians, to name more significant threats posed. The magnitude of the problem and the seriousness of the threat is not evidenced by any consistent or reliable statistics about the actual size of the ‘foreign crime problem’ in Canada. Often the subject is raised with the official disclaimer, ‘we’re not talking about big numbers but...’. The image of the ‘iceberg’ used in relation to the triad threat is illustrative in this regard in that it is evocative of a massive, yet largely invisible danger. Indeed, one of the difficulties in approaching this issue is that the apparent absence of the problem is itself used as evidence that the enforcement people are doing their jobs well and that they must continue to be supported lest the thin blue line dissolves and disorder reigns.

However, from the analytical perspective developed here, the magnitude and seriousness of the ‘problem’ as it has been constituted over the 1990s is, in fact, evidenced by the extent to which it has been acted upon legally, politically and socially and the number and range of agencies, agents and resources that have been enlisted in the ‘fight’ against it. From this perspective, there is little question that the ‘problem’ of criminality has become the central preoccupation and justification of exclusionary immigration policy in the 1990s. It is in this sense that this chapter has proposed that immigration law and policy is today a central and largely taken for granted mechanism of national crime control; that exclusionary immigration in the 1990s is ‘governed through crime’.

One further distinction between immigration enforcement and criminal justice enforcement needs to be addressed. Criminal justice enforcement is guided by the socially, politically and legally sanctioned objective of punishment. Criminals are caught and punished. Immigration enforcement on the other hand has nothing officially to do with punishment. As discussed in Chapter Two, detention and deportation and related enforcement processes and practices are constructed as straightforward administrative measures underpinned by notions of state sovereignty and territorial rights. Thus, while
immigration enforcement may look like criminal justice enforcement and while it may be experienced as punishment, it is neither. At least not in its representations. The following chapter thesis turns to a detailed examination of just what immigration enforcement looks like and just how it is experienced in the particular context of detention.
Chapter Eight

Detention and Deportation:
Sovereign Power, Carceral Conditions and Penal Practices

*Arrest is the Political Art of Individualizing Disorder*¹

I) Sovereign Penality: The Death of Michael Akhimen in Immigration Detention

On December 17 1995, 39 year old Michael Akhimen’s lifeless, naked body was found in a bathtub filled with water; he had slipped into a coma from which he did not awake. Akhimen was a Nigerian refugee seeking protection in Canada. He died in the ‘Celebrity Inn’ Immigration Holding Centre.² Shortly before his death, Akhimen had written the following words to Immigration officials: “[I would rather] die in Nigeria for a reason than waste away in [detention in Canada] when I had done nothing wrong.”³

Michael Akhimen had become increasingly unwell while in immigration detention. His pleas for medical attention were met with derision and disbelief by the facility’s security guards. On several occasions he had been placed in solitary confinement after verbal confrontations with the guards. Twice before his death, Akhimen had been found unconscious by security guards in this segregated room.⁴

While in detention, Akhimen suffered from nausea, dizziness, fuzzy vision and fatigue. He lost weight and was having difficulty eating solid foods. Akhimen made 12

---


²Lawyer Chile Eboe Osuji from the Nigerian Canadian Association represented Akhimen at the Coroner’s Inquest. He presented these details of Akhimen’s case in his deputation on detention for members of the Inter-American Commission on Human Rights in Toronto: October 22, 1997

³*NOW* “Guards Ignore a Dying Man’s Cry for Help” Vol.15, No.43. June 27-July 3, 1996

⁴Ibid..
written requests for medical treatment, and many verbal ones.\(^5\) He even offered to pay for a check-up with his own money. He was either ignored or disbelieved. On the one occasion that he was seen by the doctor at the centre, the doctor had found nothing to be seriously wrong with him, although no tests were carried out to confirm this conclusion.\(^6\)

After his death, it was determined that Akhimen had been suffering from a diabetic disease (‘diabetic keto acidosis’).\(^7\)

In frustration and desperation, Akhimen indicated that he would rather return to Nigeria than remain in detention in Canada and he withdrew his refugee application. His return was further delayed, and his detention prolonged, because CIC had lost his birth certificate, seized from him upon his detention.\(^8\)

On December 17th, Akhimen asked security for some water. When they did not bring him any, Akhimen went without security escort to the kitchen to get it himself. Upon finding him outside of his room and without security escort, guards forcibly returned him to solitary confinement and left him there.\(^9\) Shortly thereafter Michael Akhimen, alone and still in solitary confinement, slipped into a coma and died.

Michael Akhimen had sought protection in Canada in accordance with national and international human rights law. His claim to be a ‘deserving victim’ was doubted by immigration officials. As a direct result he was subjected to coercive sovereign power and penalty: power which, through detention and deportation, restraints, subjects and ultimately expels the very bodies of sovereign transgressors. This sovereign penalty, which operates in the context of exclusionary immigration law and policy continues to be distinguished from punishment; a distinction which serves to protect its coercive and\(^{\text{\footnotesize ---}}\)

\(^5\) Osuji, deputation on detention for the Inter-American Commission on Human Rights. Toronto: October 22, 1997

\(^6\) NOW “Guards Ignore a Dying Man’s Cry for Help” June 27-July 3, 1996

\(^7\) Amnesty International Human Rights Report 1997

\(^8\) NOW “Guards Ignore a Dying Man’s Cry for Help” June 27-July 3, 1996

\(^9\) Osuji, deputation on detention, October 22, 1997
punitive operations from serious judicial, political and public scrutiny.

Michael Akhimen’s death was the subject of a coroner’s inquest. It ruled that Michael Akhimen died of natural causes and did not address the harm of solitary confinement. It held that it was not its place to comment on immigration policy and practices and therefore ignored the state sanctioned punitive and coercive circumstances and practices which had, at the very least, exacerbated Akhimen’s suffering before he died and at worse had actually contributed to his death.\[10\]

If judged undesirable for reasons of criminality, non-citizens are held in provincial and federal correctional institutions. If criminality is specifically not an issue, non-citizens are detained in immigration holding centres. In the United States, as in Canada, airport motels have been used as holding centres. Michael Akhimen died in the ‘Celebrity Budget Inn’ in Mississauga. That sovereign power operates today in motels and not in the town square reflects the degree to which the coercive edge of sovereign power has been reconfigured in accordance with the discursive influence of liberal humanitarian and legal influences. Indeed, there is little doubt that had Akhimen’s final days been a matter of public spectacle, the public would have opposed it, at least there certainly would have been no cheering. However, the extent to which the ‘Kafka’ hotels have been modified and operate in accordance with ‘medium’ security carceral conditions speaks to the limits of these mitigating discourses in altering the fundamentally punitive and coercive nature of the sanction imposed. Indeed, the relatively hidden nature of sovereign penalty in this area contributes to and exacerbates the continuing punitive and coercive conditions of immigration detention.

\[10\]NOW “Guards Ignore a Dying Man’s Cry for Help” June 27-July 3, 1996
II) The ‘Celebrity Inn’ Immigration Holding Centre

"Travellers Under Cloud Stay at the Inn of Unhappiness"

Welcome to the Celebrity Inn - an Immigration Holding Centre, in bureaucratic parlance. In reality, it is an immigration jail for the unlucky, the fraudulent and the suspect.\(^{11}\)

In the United States, hotels which double as immigration holding centres are commonly referred to as "Kafka motels"\(^{12}\) The designation is evocative. Perhaps the most enduring feature of Franz Kafka’s writing is his powerful depiction of the absurdly coercive qualities of modern bureaucracy. The main character of his novel, The Trial, is quite literally a ‘prisoner of administration’. He is unable to learn anything about the charges against him and his very life and liberty is in the hands of faceless and anonymous accusers. Significantly, he is known only as ‘K.’, his identity irrelevant in the maze of mysterious and coercive workings of a strange punitive administrative system. While the workings of this system are mysterious and incoherent, its coercive power is real and its punitive impact profound.

The ‘Celebrity Inn’ in Mississauga is Ontario’s own Kafka motel. ‘Celebrity’, the immigration detention centre is located in one wing of the fully functioning and busy airport motel, the Celebrity Budget Inn. Conveniently located less than a kilometre from Toronto’s Pearson International Airport, Celebrity Inn promises to “treat all guests as celebrities”. Kafka couldn’t have said it better.

There are two entrances for those who stay at the Celebrity Inn, one for those who deserve to be welcomed and one for those who do not, one for those who can leave freely and one for those who cannot, one for those who are treated like criminals and one for those who are not. At the Celebrity Inn the included and the excluded bunk in different

---

\(^{11}\)Globe and Mail. December 28, 1995

wings of the same motel. This illustrates quite dramatically the contrasting destinies of those deemed through the operation of discretionary immigration powers to be deserving and/or desirable and those deemed otherwise.

As already mentioned, individuals whose cases involve criminality are not sent to Celebrity but to provincial criminal justice ‘correctional’ facilities. This makes detention at Celebrity a particularly interesting and analytically revealing case. CIC’s ‘holding’ centres are specifically designed to accommodate non-criminal. Thus while one might expect that those held on immigration hold in more secure correctional institutions for reasons of criminality would likely be subjected to the coercive and punitive dimensions of criminal justice incarceration. ‘temporary holding centres’ which are specifically for detaining non-criminal immigration cases might reasonably be expected to operate under a different, non-carceral, regime. However notwithstanding the absence of ‘criminals’. and notwithstanding official declarations to the contrary, immigration detention - even in a ‘holding centre’ for non-criminals located in a motel - is a distinctly carceral and penal experience.

The distinctions between immigration detention and deportation and criminal justice incarceration hinge centrally on the concept of punishment: immigration ‘administers’ offenders whereas criminal justice punishes them.\(^\text{13}\) It is however telling that despite this longstanding legal distinction, many people, including the Minister of Immigration\(^\text{14}\) and one of the members of the Immigration Legislative Advisory Group, are a little unclear on this point. Indeed even ‘experts’ confuse the objectives of immigration detention and criminal justice incarceration; “...we use detention punitively but we don’t use it productively...Its counter-productive, and punitive is only one

\(^{13}\)The legal distinction between immigration detention and deportation and punishment is discussed in Chapter 2.

\(^{14}\) As mentioned in Chapter 2, Minister of Immigration, Lucienne Robillard. stated explicitly during the 1997-98 CIC Standing Committee hearings that punishment was but one of the purposes of detention.
Clearly, the emergence of immigration law and policy as a major mechanism of crime control, the related proliferation of crime control enforcement powers, practices and attitudes, the increasing currency of deterrence rationales in this field and the criminalization of new immigrants and refugees have all contributed to the blurring of this abstract legal distinction. This distinction disintegrates further when one considers the actual carceral conditions and penal practices of immigration detention and deportation.

i) Carceral Conditions

The presence of a medium security detention facility in the isolated rear wing of the Celebrity Budget Inn would likely come as a surprise to most of Celebrity’s paying customers. From the main entrance to the Inn, there are few visual clues that it doubles as a medium security detention centre. Around the side of the building, a keen observer might notice a surveillance camera mounted on the outside wall, just above a steel door entrance with a coded locking mechanism just beside it. This is the visitor’s entrance to the detention facility. It is permanently locked. To enter, visitors ring a buzzer and their image is recorded and transmitted to a security officer within who may then deactivate the lock and allow entry.

The door opens into a large, dingy, yellowish room which is empty of furniture except for the chairs which line the walls. There is another surveillance camera and a pay phone in this room. Apart from posted notices which detail certain institutional regulations regarding visitation, there is no other printed public information available here. One corner of this room has been sectioned off and a security guard permanently posted behind fortified plexiglass acts as the facility’s ‘visits’ officer’. Just behind this security post is the ‘detainee visiting area’ in which detainees sit at cubicles and meet their visitors through plexiglass barriers and communicate through a telephone. There are

---

15 Susan Davis, Member of ILRAG, CIC Standing Committee Hearings. February 25, 1998:8
twelve visiting 'stations'.

In order to enter the 'inner' regions of the centre, the locks on two more steel doors need to be deactivated by security. A metal detector lines the frame of one of these doors. Immediately on the other side are blue tiled stairs which lead up to the second floor of the 'detention wing' of the Inn. It is indeed a very gloomy place: dim lighting, non-descript beige/brown walls, long empty corridors. It is exceptionally clean: not new, nor necessarily in good repair, but clean. Celebrity, the detention centre, is serviced by the cleaning staff of Celebrity, the hotel. The detainees rooms and every floor of every room is cleaned daily, in the winters sometimes twice or three times because of the snow and salt from outside. The disinfectant smell of ammonia and cleaning fluids hovers throughout. The walls, inside and out, are frequently painted. The air quality within the facility can only be described as terrible due to the permanently sealed windows and lack of ventilation and fresh air circulation.

Located at the top of the immaculate but gloomy stairway are the separate entrances to the male and female ‘dining’ and smoking rooms and the cafeteria style kitchen. A long, dimly lit motel hallway leads away from the dining rooms to the detainees rooms. The sex segregated dining rooms are also used during the day for scheduled ‘telephone time’. The men’s area has 7 free phones for local calls and 2 pay phones, the women’s has 4 free phones and 1 pay phone. There are security posts at each end of the hallway, and at each entrance of the dining areas. There are surveillance cameras in the hallway and in each of the ‘common’ areas. Both dining areas are brightly lit by natural light; windows sealed with ‘lexon’ (a high quality plexiglass) line their outside walls. The windows in the men’s area look out onto Airport Road and the Airport

---

16The air quality at Celebrity has been frequently raised as a matter of concern. In 1997, the Inter-Church Committee on Refugees issued a report after being given a tour of Celebrity. The report raised many items of concern relating to health and safety standards at the facility, including that of poor air quality. Legal support for criticisms of immigration detention is commonly sought in International Treaty Standards, for example the poor air quality at Celebrity contravenes the UN Standard Minimum Rules for the Treatment of Prisoners. (See fn.47)
Cargo bay. The men's dining area is considerably larger than the women's. In both, long cafeteria style tables occupy most of the floor space. A bookshelf in the corner of each room (the "library") holds dog-eared, second hand and multi-lingual books donated by non-governmental organizations. Posted notices in each room provide a few phone numbers of frequently needed agency contacts (legal aid, a couple of metro shelters, embassy numbers). detail emergency procedures (fire) and other regulations (eg. "luggage. money may only be accessed between 1800 to 1900 hours. 6-7 pm unless otherwise specified"; nurse and doctors hours. visiting hours. phone messages. CIC 'Rules for Detainees'). and provide a daily listing of court dates.

The immigration officials who run the facility. Enforcement Detention Officers (EDOs). and their support staff, work primarily out of three rooms. the entrance to which is located kitty corner from the dining areas. A security officer is posted outside of their door. which is usually closed. It is here that the administrative work of the facility is done. The resemblance of the role of Celebrity's EDOs to that of correctional officers is striking. EDOs wields an enormous amount of discretionary power over detention and detainees. including the power to overturn an enforcement decision to detain and replace it with their own decision to release within the first 48 hours of detention. In addition to admission and release powers. the EDOs at Celebrity are responsible for every aspect of the daily administration of the facility; from managing contracts with private suppliers (security\textsuperscript{17}. hotel\textsuperscript{18}. medical\textsuperscript{19}) to hearing and investigating complaints to taking

\textsuperscript{17}CIC contracts a private security company to guard the detainees. The current contract is held by Wackenhut Security Inc., an American company which specializes in prison security.\textsuperscript{18}

\textsuperscript{18}In 1997. CIC signed another three year contract with the proprietor of the Celebrity Budget Inn. Since January 2000. the contract is renewable every 6 months. Plans are in the works for the purchase by CIC of a new. centralized detention centre in Ontario which would be entirely administered by CIC. For this reason. in January 2000. CIC has moved to six month renewable contracts with the Celebrity Budget Inn. (Interview with SIO at Celebrity, February 3, 2000)

\textsuperscript{19}CIC contracts the part-time services of a community based doctor and 2 nurses. (Ibid..)
disciplinary action against difficult detainees. In these three rooms, phone calls are made, interviews carried out, reports completed, statistics generated and photocopies made. In 1997, a rather eerie representation of the increasingly equivalent roles and powers of immigration and law "enforcement" was located in CIC's photocopy, fax and supply room - a toy gun and handcuffs sealed in clear hard plastic with the direction to "break plastic in case of emergency."20

Across from the Immigration office is the "children's playroom". This room was created in response to pressure from the Toronto Refugee Affairs Council (TRAC) concerned about the frequent presence of young children in Celebrity and the absence of any special programs to address their particular needs. The children's playroom is empty of furniture. miscellaneous toys litter the floor: a plastic slide. a rocking horse. a 'make-up' station. It is not the practice to detain young children. They are there with their detained parent(s) as 'guests' of Immigration.21

The remaining rooms down the second floor hallway are the women's accommodations. The men's rooms are in a separate wing of the Inn. There used to be several 'family' rooms in the women's wing where families were permitted to remain together but this practice was eliminated when "a few poorly behaved detainees spoiled it for everyone." Families may still eat together and spend time together in a designated 'family room'. but husbands/fathers now all sleep in the men's wing and children remain

---

20 As observed during my field work at the 'Celebrity' detention centre between August 1997 and June 1998. During this period, I volunteered as the on site case worker for the Toronto Refugee Affairs Council (TRAC). TRAC has an agreement with CIC which provides office space and telephone access for a case worker who provides information and referral services to detainees. The case worker is specifically prohibited from performing advocacy services, they are only there to 'facilitate access to the system'. I performed this function for TRAC, two days a week for 9 months. I am particularly grateful to Fred Franklin of TRAC for having facilitated and supported my work at Celebrity, and to the CIC officials at Celebrity who for the most part responded to my queries and tolerated my curiosity while I was there, and who were open and forthcoming subsequent interviews. See footnote 25 for additional information regarding TRAC.

21 Interview with SIO at Celebrity. Feb.3, 2000
with their mothers in the women's wing. All rooms have two single beds, a dresser, a
cable television and an off-suite bathroom. While there are glass mirrors on the walls, and
there are some light bulbs in fixtures, bars house the non-functioning airconditioners and
the doors have been removed from the room's closets.

Escape, not violence, is the primary security concern. Each room has an outside
window, which is sealed with reinforced plexiglass. Beds have also been modified: the
ordinary, iron rail bed frames were replaced because "people were ripping them out, using
them as pry bars and weapons, so we got rid of that."

On the first floor of the detention wing, the rooms which line the hallway are used
for a variety of purposes: several are designated as 'meeting rooms' for detainees and
their lawyers or other 'professional' contacts; in 1999 several were adapted to function as
'video-conferencing' detention review rooms where detainees now present their case for
release, via video, to Immigration adjudicators. Until recently, detainees were transported
off site to attend these reviews. They now are done from Celebrity with the technology of
video-conferencing. This manner of conducting the reviews has been criticised for further
undermining a detainee's ability to make their case persuasively. Eusavio Garcia, part
time case worker for TRAC at Celebrity and Refugee and Settlement Worker for the
Quaker Committee for Refugees, observed that because so much of the detention release
decision depends on the perceived 'credibility' of the detainee, the disconnected and
impersonal medium now employed presents a further obstacle to detainees. Another
consequence of this initiative is that because lawyers and other representatives of the
detainees do not have to be physically with their client during the proceedings, many are
choosing not to make the trek out to the airport location of Celebrity. This separation of
client and counsel presents a further disadvantage of this practice.

---

22 Ibid..
23 Ibid..
24 See chapter two for a discussion of these detention reviews. Interview with Eusavio
Garcia, part-time TRAC case worker, February 3, 2000
In addition to the video-conferencing rooms, the facility's doctor and nurses administer health care services out of two rooms; there is a detainees' 'baggage' room: a room for the Toronto Refugee Affairs Council (TRAC) caseworker; and finally there are several rooms which are used for the purposes of segregation and solitary confinement for health or security/disciplinary purposes. It was in one of these rooms that Michael Akhimen died.

The beige and brown hallway on the first floor of the detention wing leads ultimately to the security headquarters of the facility. Three rooms, the main security (or 'supervisor's) office, the 'admissions and discharge' office, and the 'holding room', form a triangle at the end of the hall. The Supervisor of Security and the Head Guard are permanently posted at the main security desk. The supervisor's desk faces a wall of television monitors which are continuously transmitting the surveillance images generated by the facility's five internal and seven external surveillance cameras.

Across the hall from the security office, is the detainee 'holding room'. Of all the rooms in the Inn, this room most closely resembles and evokes the popular image of a police cell. It is small, no more than eight to ten feet square. It is a dirty shade of pale, institutional yellow. It is lit by fluorescent lights. There are no windows. Wooden benches are bolted to the walls. There is a good size plexiglass window in the door to permit viewing of the room from the outside when the door is closed. Once closed, the door

---

25 The Toronto Refugee Affairs Council is a non-governmental group that has maintained a presence at 'Celebrity' since 1985. TRAC caseworkers posted at the detention centre work interview detainees, assess their needs, provide information and make appropriate referrals to community and legal contacts. TRAC also holds 'detention review seminars' to provide information and guidance relating to the detention review process. CIC reports being pleased to have TRAC on site and leaves all programming and special service delivery to TRAC's volunteers, for example English as a Second Language, Spiritual Counselling, Art Therapy, Assistance in filling out official forms. Sadly, in February 2000, these initiatives had fallen to the wayside due to TRAC's severe resource limitations. The implications of this are serious, as put by the Celebrity SIO, "If TRAC doesn't do it, then it isn't done." (Interview February 3, 2000)

26 Interview with SIO at Celebrity February 3, 2000
cannot be opened from the inside.

The third room in the security triangle is the Admissions and Discharge office. The safety deposit boxes for the valuables and money of the detainees is located in this room. Just beyond these 'rooms is the third entrance of the Inn, the detainees’ entrance. This opens into the 'loading and unloading' area of the compound which forms a corner of the detainees' exercise yard. The ‘loading and unloading’ area is just inside the constantly monitored double gate through which detainees are delivered to or transported from the facility in customized, dark blue Immigration Enforcement vans. After being ‘unloaded’, detainees are delivered under escort through the detainees’ entrance to the main security desk and the admissions and discharge security officer.

The ‘exercise yard’ used to be the parking lot for the motel rooms in this wing. It measures about 25 feet wide by 100 feet long. It is empty save for two picnic benches which are bolted to the pavement and a poorly situated basketball hoop. It is surveilled by several cameras mounted to the outside walls of the Inn. Cigarette butts quickly accumulate on the cracked asphalt. Other than the basketball hoop and the occasional soccer ball, the only provisions for those detained to exercise are a stair-master and an fitness bike located in the children’s playroom.

The ‘exercise yard’ and the loading and unloading area are encircled by two wire fences. The outside fence is twelve feet high and is capped with barbed wire; of the ‘ordinary’ as opposed to the ‘razor’ variety. The inside fence is eight feet high with ‘...an inward leaning overhang which is covered with mesh so that you can’t get your fingers through it and climb up it.’ This recently added innovation was a response to the proven ability of a certain number of detainees to climb up and over the previously existing single barbed wire fence with no mesh overhang; ‘...we’d have these athletic young guys who could get over anything.’ Reportedly many did. However, since the modification, nobody has gone over.28

---

27Ibid..

28Ibid..
CIC contracts a private security company to police Celebrity. The current 2 year contract is with the Wackenhut Security Company, an American company that specializes in prison security. During the day, there are 14 guards on site, including 3 drivers, the Security Supervisor, the Head Guard, the Visits Guard, and the Admissions and Discharge Guard. During the evening and midnight shifts, the number of guards on site is reduced to 9 or 10. CIC provides Wackenhut guards with a 5 day specialized training session. Guards, who are paid about $9.50 an hour, must also pass a minimally demanding physical fitness test. They receive basic defence training, training on the proper use of body restraints and training on the security Post Orders. While guards are instructed to respect the principle of non-discrimination, they do not receive any form of cultural sensitivity training from CIC.

Section 2.06 of the 1996 Post Orders summarises the responsibilities of the security contractor:

- Supervise persons being detained.
- Admit/release detainees from the Holding Centre.
- Provide information to new detainees concerning rules of the Centre.
- Take control and be responsible for the personal effects of detainees.
- Order meals and verify delivery.
- Conduct frequent and unscheduled room checks.
- Admit visitors and ensure that visitors are not in possession of weapons, alcohol etc.
- Obtain medical treatment for detainees as required.
- Evacuate the Centre in the event of an emergency.
- Operate metal detecting devices.
- Complete reports.
- Transport all persons under order of detention as requested by Citizenship and Immigration.
- Prevent escapes.
- Apply restraints.

---

29 Interview with SIO at Celebrity, February 3, 2000

30 CIC. Detention and Removals, Mississauga Immigration Holding Centre Post Orders. February 29, 1996.
ii) Penal Practices: The Coercive Subjection and Control of Unruly Bodies

The technologies and practices employed by CIC and the private security company in the running (policing) of immigration detention are those conventionally associated with, and historically employed in, criminal justice punishment. That is to say, the ‘how’ (the means, the strategies, the practices, the ‘technologies’ of governing) of both criminal justice imprisonment and immigration detention are, to a significant degree, the same. However the underlying rationales for the employment of these coercive practices in the fields of criminal justice and immigration are vastly different. The ‘humanizing’ influences which in the nineteenth century reformed, at least to a degree, the prison system in order to transform the character of prisoners left untouched and unchanged the application of the essential sovereign nature of immigration penalty. While liberal legality has influenced the development of policy and practice in the coercive governance of immigration detainees and deportees and the manner and ‘form’ in which penal technologies and practices are employed, the object of the actual penal practices employed is to achieve bodily subjection, control and ultimately expulsion: no more, no less.

In the Celebrity detention centre, these procedures and practices are governed by the above cited Post Orders. The ‘Post Orders’ detail CIC’s operational procedures relating to the security aspects of immigration detention. CIC admits the term ‘post order’ is a misnomer as it implies a public posting of the ‘orders’ which govern the facility. In fact, they are more accurately described as the Department’s instructions to the Security personnel that police Celebrity. In true Kafka fashion, these ‘Post-Orders’ are not in fact publicly ‘posted’ and accessible to all, but rather are chained to the security posts in the facility and are only accessible to Security. In February 1996, under some pressure from the public criticism and scrutiny that was sparked by Michael Akhimen’s death, the Department transformed previously looseleaf, largely ad hoc departmental memorandums and directives regarding security matters into a sirloin bound mimeo publication. The

---

31 Interview with SIO at Celebrity, February 23, 2000
1996 *Post Orders* responded to the recommendations made by the Coroner, by detailing and strengthening the sections relating to the provision of health care within the facility.

**a) Cuffs and Shackles**

*Handcuffs are revealed not merely as the technical instruments of security...Requiring a deportee to travel in handcuffs confirms that it is the body which is at stake here, the unruly body identified as the illegal immigrant.*

The use of body restraints on detainees is standard policy and practice during transport to and from the Celebrity detention facility. In particular, handcuffs and leg irons or shackles constitute the "standard" from which deviations may arise. Exceptions may be made with the preauthorization of the Enforcement Detention Officer (EDO) or the Security Supervisor. In general, reduction of the standard uses of body restraints may be made in the case of "children, old people, very pregnant women, disabled persons or person's who should not be handcuffed for a specific reason."

The standard must be applied in full in the case of male detainees being transferred to or from a more secure correctional facility (jail). In such cases, handcuffs and leg irons may be further supplemented by the use of a "transportation belt" if the guards judge that there is a risk of violence. A transportation belt works to secure handcuffed wrists at the waist level thereby eliminating the danger of the person using their cuffed wrists as weapons. Exceptions to this standard must be approved by a Senior Immigration Officer. Female detainees being transported from a correctional facility are also always cuffed but are spared the leg irons unless specifically directed by the EDO or the Security Supervisor. The use of these traditional penal instruments of bodily control

---

32 *Post Orders* (‘Restraints’ s.12 and ‘Transportation’ s.11)

33 Alison Young *Imagining Crime* London: Sage, 1996:71

34 Interview with SIO at Celebrity, February 3, 2000

35 The relevant sections of the *Post Orders* for these regulations are: ss.11.08, 11.13 and 11.20
is guided in the context of immigration detention by the principle of 'safe restraint' and is justified by reference to the need to protect safety and prevent escape.

Still, handcuffs, leg irons and transportation belts cannot help but evoke images of criminal justice penalty and the forcible subjection and confinement of suspected criminal offenders. The physicality and brute forcefulness of these instruments of restraint are also powerfully reminiscent of medieval penalty, particularly when linked with the expulsion. Handcuffs and shackles are standard fare for deportees as well whose bodies are forcibly and physically subjected as part of the expulsion process. However, their power is not merely negative and repressive. to the contrary it is a 'productive force'. As Allison Young comments, the use of body restraints on detainees and deportees "is part of a programme...to produce docile subjects...to train the individual body, the social body and the nation." 37

This association of forced bodily restraint and expulsion with pre-modern and early modern sovereign modes of penalty (stocks and banishment) is heightened by tales of abuses and extreme bodily interventions which occasionally come to light. In 1990, there were reports of 'unco-operative' deportees who were forcibly drugged and sedated during their removal. 38 Enforcement officers were also known to seal the mouths of particularly unruly and vociferous deportees with duct tape. 39 In 1999, 5 'manacled and shackled' Nigerians were deported from Canada aboard a US Federal Service flight nicknamed ‘Con Air’. 40 CIC paid $4,800.00 (American) for each of the one way tickets. reportedly only a third of the cost of using commercial airlines. 41 The U.S. leases planes

---

36 Interview with Immigration officer at Celebrity, February 3. 2000.
39 Ibid.
40 Ibid.
41 Ibid.
which are used in a fleet officially known as the “Justice Prisoner and Alien Transportation System”. The flights carrying ‘aliens and criminals’ are heavily guarded by armed U.S. Marshals:

Con Air is an Airline with a Difference. The standard operating procedure by the flight attendants - armed US Marshals clad in bullet proof vests - is to handcuff obstreperous passengers to their seats. Really disruptive ones have their handcuffed hands taped around tennis balls.⁴²

On their flight, the Nigerians were also ‘escorted’ (subdued, transported and guarded) by CIC Escort officers. While the use of ‘Con Air’ to deport the Nigerians was carried out ‘in utmost secrecy’. CIC officials later declared that it represented an “innovative method” of deporting “disruptive people” in an “economic and efficient” way. As evidence of the disruptiveess of the Nigerians in question, an immigration source stated that the Nigerians “...were kicking, scratching, fighting and spitting as they were being boarded.” All 58 passengers on this flight were described as criminals. According to CIC, the Nigerians “...had committed crimes ranging from theft and drug offences to assault and carrying a concealed weapon.”

The symbolic importance of the use of restraints on deportees is perhaps more forcefully illustrated by the following example recounted during the 1997-98 Standing Committee hearings:

The newspapers have reported actual cases where persons were drugged before being sent back without any medical supervision, or handcuffed for their return flight, or even put in leg irons....It is hard to forget the case of a young Dominican who was deported a short while ago from Canada. Despite the fact that both of his feet had been amputated, he was nevertheless handcuffed.⁴³

The symbolic and historical association of coercive bodily restraints with criminal. penal practices is deeply embedded and internationally understood. Criminals

⁴²Ibid.

⁴³Jean-Michel Montbriand, President, Association Quebecoise des Avocats et Avocates en Droit de L’Immigration, CIC Standing Committee Hearings. April 23, 1998:3
are handcuffed. Really dangerous criminals might be shackled and cuffed and belted. The people being transported to and from Celebrity are neither, yet in this important bodily respect they are acted upon as if they were criminals. State sanctioned, direct and physical subjection and control of bodies by force through the use of body restraints is universally associated with criminal justice enforcement. Their use in the non-criminal, administrative context of immigration enforcement both evidences and reproduces the association between immigration and crime control, immigrants and crime and the coercive edge of traditionally sovereign power which characterises 'immigration penalty'.

This point was voiced succinctly by Louise Hardy, Member of Parliament for the Yukon, during the 1997-98 Citizenship and Immigration Standing Committee Hearings:

The fact is....that they do shackles them, and they shackle them all the time...If there's no suspicion of them being criminals, I question why they should be there in a centre. Its called a detention centre. but I've been through a lot of jails, and its like a jail and it has the same security. It has more security than the jail in the Yukon. So these people are subjected to treatment as if they were criminals, whether they are or not.

For those being physically subjected, the use of body restraints is experienced as

---

44 While beyond the scope of the current analysis, it is nonetheless important to note that international legal instruments provide legal support for many of the criticisms levelled at immigration detention conditions and practices. For example, on the subject of restraints, lawyer David Matas points out that “people who are put in detention are typically put in handcuffs, chains and irons, even though it is against the [United Nations'] Standard Minimum Rules for the Treatment of Prisoners'....”(CIC Standing Committee Hearings. March 25. 1998:13) CIC has recently acknowledged the importance of these legal standards and has incorporated them in their “National Standards and Monitoring Plan for the Regulation and Operation of CICs Detention Centres” (Draft copy. February 5. 1998). This document details all the different aspects of immigration detention and identifies the relevant international legal standards for each. These include relevant provisions of the Canadian Charter of Rights and Freedoms, UNHCR Guidelines on Detention of Asylum Seekers, the U.N. Body of Principles for the Protection of All People Under Any Form of Detention or Imprisonment and the above mentioned UN Standard Minimum Rules for the Treatment of Prisoners.

45CIC Standing Committee Hearings, April 22, 1998:11
a distinctly penal practice; coercive, punitive, humiliating and undeserved. The official policy justification for this use of restraints stresses, in part, safety and security issues. However, significantly, it also recognizes the symbolic importance of their use:

The practice is that when people are being removed or transported for whatever reason outside of the centre, usually to a hearing and sometimes to removal, they will be handcuffed. It's a recognition that the people are in detention. Its partly related to safety and security issues. We do have, from time to time, people who do attempt an escape. Sometimes that escape will involve some violence. If we're going to maintain control of the people, we do use handcuffs. We recognize that being detained, which includes the use of handcuffs in those circumstances, is a serious issue (my emphasis).

That immigration detainees are treated like criminals is clearly evidenced by the use of these conventionally penal, coercive instruments on the bodies of 'non-criminals' in the context of an enforcement oriented immigration 'penality' that is anything but a 'purely administrative' proceeding. Indeed, the toy gun and handcuffs hanging in the Immigration office at Celebrity, surely an example of rather grim gallows humour, also indicates a rather dark and cynical understanding of the shared bodily objectives of immigration and law enforcement.

---

46 'Why do they treat us like criminals?' is perhaps the question I heard most frequently during my time at Celebrity as TRAC's on-site caseworker. The use of handcuffs was a constant focus of anger, bewilderment, humiliation and frustration. Of particular concern to the detainees was the use of handcuffs on mothers in front of their children. This concern, which was also voiced during the 1997-98 hearings, has been responded to in the forthcoming CIC Post Orders, which bans this practice. (Interview with SIO at Celebrity, February 3, 2000)

47 Neil Cochrane, Director Case Presentation, Enforcement Branch, CIC. April 22, 1998:12
Coercive Regimes

a) Admission Procedures

The dark blue Immigration Enforcement van passes through the gate in the doubled, 12 foot high barbed wire fence. A surveillance camera tracks it as it enters the ‘loading and unloading’ area of the compound. The detainees being transported\(^4\) are ‘unloaded’ and escorted by guards into the facility. Body restraints are removed and detainees are placed in the ‘holding room’ where they are ‘frisked’ (‘pat’ searched). their shoes are removed and checked, and a hand-held metal detector is passed over their bodies by a guard of the same sex. Any luggage is seized, searched (in the presence of the owner) tagged and stored in the permanently locked baggage room. All valuables and any money are recorded and locked in safety deposit boxes. If any identification documents are found in the possession of detainees, regulations detail which are to be seized and to whom they are to be forwarded. Detainees may bring clothes and other ‘non-threatening’ items to their rooms. They may keep up to 30 dollars on their person for their personal use. Detainees have access to their seized luggage and valuables between 1800 and 1900, or by special authorization.

New detainees are assigned a log number, a room number and the time of their admission is recorded. Security must ensure that each of the new detainees are provided with ‘..and allowed to read/view the rules of the Centre.’\(^5\) Detainees are provided with two information handouts: ‘Rules for Detainees’ and ‘Information for Person’s Detained’. These do not contain the same information as the Departmental ‘Post Orders’. nor are they summaries of the Post Orders. The ‘rules’ provided to detainees are largely prohibitive and govern their behaviour while in detention. Items covered in the rules include the seizure of certain belongings, restrictions on mobility and movement, visiting hours, doctors hours, television volume controls, the expectation that detainees keep their

---

\(^4\)Admission procedures are set out in s.4 of the Post Orders

\(^5\)The procedures relating to the transportation of detainees are found in s.11 of the Post Orders
rooms “neat and clean”, lights out at 11:30pm, the presence of TRAC on site and the complaints process. It is important to emphasize that while Celebrity is not a jail, jail is never very far away for those in detention. Indeed, the spectre and ever-present threat of jail is a key and powerful technology of control at Celebrity: so powerful that it is explicitly detailed in the final and penultimate rule:

#14. These rules are designed for the safety, security and comfort of all persons in the Centre. You are expected to respect them. Disruptive behaviour, including damage of property, will not be tolerated. Such behaviour may lead to your transfer to a more secure detention facility.  

The ‘information’ provided to detainees outlines briefly the detention and release process and their rights under the law.

Before being escorted to their rooms, new detainees must be frisk-searched one more time. From this point on, detainees may not leave their room except under security escort and only for certain regularly scheduled and carefully coordinated and controlled movements and activities (meals, ‘fresh air’, ‘telephone time’, visitor’s hours). or with permission for special requests (visit with the doctor or a lawyer for example). Visits with other detainees in their rooms are prohibited.  

b) Restrictions on Movement

The movement of detainees within the facility is carefully coordinated and controlled by security. Detainees must have a security officer escort when being moved from one location of the facility to another. They must have special permission to leave their rooms at unscheduled times and if permission is granted, they must be escorted by a guard at all times. If a guard is not available for escort at the time of the request, the detainee is locked in the holding room until a guard becomes available. A minimum of

---

50 CIC “Rules for Detainees” Immigration Holding Centre (Mississauga)
51 Ibid.
52 Post Orders, ss 8.00
one officer must escort a maximum of 5 detainees (the security officer must walk behind the group). More than 5 detainees must be escorted by 2 security officers (one in front and one behind) and "under no circumstances are there to be more than 10 detainees moved at one time." For 'mass movements' of detainees, such as that required at meal time, telephone time and fresh air time, 10 groups of detainees are moved at a time from security post to security post. The security procedures governing the movement of detainees and their supervision at meal times and fresh air time are particularly detailed and specific.

c) Patrols, Checks and Searches

Floor patrols, outside patrols, room checks and room searches are carried out by security personnel on a regular basis. Floor Patrols must take place 'on an unscheduled basis', at least once every half hour. Room checks are also performed regularly. The specific provisions which detail the manner in which they should be carried out at night reveal there two functions: to confirm the detainee is both physically present and alive: "...shine a flashlight on the body of the detainee, to ensure that there is breathing movement." The section specifically warns against shining the light in the eyes of

- Post Orders s.8.04-8.09 ("Banquet Hall") and s.8.10-8.19 ("Exercise Yard")

Other aspects of detainees behaviour are also regulated; for example 8-10 detainees must be seated at each table at meal times, no smoking is allowed during meal times, even in the smoking rooms, and detainees are not allowed to bring food back to their rooms. Detainees are not allowed to "touch the TV" in the dining rooms, channel selection and volume is "...under the control of the head guard." The TV will be "OFF" during meal times. The Security Supervisor has the discretion to allow detainees to remain in the dining areas to "watch something special".

- These items are covered under s. 9 of the Post Orders

Many of the sections of the Post Orders reflect the reactive process of rule-making at Celebrity: sections are added as problems requiring attention arise. The prohibition against shining a flashlight into the eyes of sleeping detainees is an example of this as is the rule governing the closure of blinds in the rooms of detainees. The windows of the men's and women's rooms face each other. There had been reports of rude and offensive
sleeping detainees. Outside patrols are conducted twice on the afternoon shift and four times on the midnight shift and guards are instructed to be on the lookout for "suspicious persons, vehicles, property damage, noise, etc." "Thorough" room searches are to be carried out at least three times a week, while the detainees are in the dining rooms. The searches are done by two security guards who are to record and/or seize anything "unusual." 56

d) Use of Force

CIC staff and security personnel at Celebrity are governed by the relevant provisions of the Criminal Code of Canada regarding their use of force. They are empowered to use "necessary force" to prevent an escape and/or to protect themselves against an assault. They are prohibited from using "excessive" force, that is, force that is intended or is likely to cause "death or grievous bodily harm". 57

e) Punitive Measures

The policing and control of detainees is also effected through disciplinary measures taken against disruptive or otherwise unruly detainees. When presented with a relatively minor disciplinary problem, the preferred initial action by the EDO is "counselling" ("talking to the guy and finding out what happened, what the problem is and trying to resolve it") 58. If that doesn't work, the EDO may confine the person in question to one of the "isolation/overflow" rooms on the first floor for an unspecified length of time by the male detainees in full view of the female detainees. A rule mandating blind closure was subsequently adopted.

56 Post Orders s.16.01

57 The relevant sections of the Canadian Criminal Code are set out in s.17 of the Post Orders: s.25 (Protection of Persons Acting Under Authority); s.26 (Excessive Force); s.34 (Self Defence Against Unprovoked Attack); s.37(1) (Preventing Assault) and s.17.05 (Rescue or Permitting Escape).

58 Interview with SIO at Celebrity, February 3, 2000
time. While CIC is quick to point out that these rooms are identical to the accommodations on the second floor and that their doors are not locked, the fact remains that those confined in these rooms are confined alone and may not leave them.

The ultimate, and arguably the most powerful sanction available to the EDO with respect to a 'disciplinary problem' at Celebrity is to transfer the person to jail. The threat of a sanction - isolation or transfer to jail - is frequently employed to control unruly subjects. Actual transfer to a prison is generally assured in the case of a more serious 'disciplinary problem' such as posed by a violent or abusive detainee or someone who has attempted escape. In 1999, several people who were detained at Celebrity went on a hunger strike. They too were transferred to prison.59

In a rather strange and symbolically evocative arrangement, the isolation/overflow rooms on the main floor are used for both punitive/disciplinary reasons (solitary confinement) and for reasons relating to the health of detainees (quarantine). It is equally grim that the sanction of transfer to prison may similarly be imposed for either punitive/disciplinary reasons or for 'health and safety' reasons (detainees thought to pose a 'suicide risk' are transferred to prison).

iv) Daily Routines and Schedules60
The daily lives of detainees at Celebrity are organized by and revolve around various 'schedules': the daily time schedule, the visiting schedule, the baggage access schedule, the telephone time schedule, the medical services schedule.

a) Time Schedule
0700-0800 Breakfast
0800-0900 Detainees to be in their rooms
0900-1100 Telephone time and fresh air (weather dependent)

59Interview with Eusavio Garcia. TRAC Caseworker at Celebrity. February 3, 2000

60Post Orders. s.8.20
1100-1300 Lunch
1300-1400 Detainees to be in their rooms
1400-1500 Telephone Time
1500-1700 Detainees to be in their rooms
1700-1800 Supper
1800-2100 Detainees to be in their room and fresh air
2100-2200 Telephone time
2200 Detainees to their rooms
23:30 T.V.s off

b) Visitor's Schedule 61

Visitation at the Celebrity detention centre is strictly regulated. Visits are permitted only between the hours of 9:30-11:15 am, 13:30 - 15:15 pm and 19:00 - 20:45 pm. In addition to the temporal restrictions, an individual detainee may not be visited by more than two visitors at a time, and may have no more than 3 visitors during any one of the three time slots. A maximum of 12 detainees may meet with visitors at any one time, a quite severe restriction as there are often more than 80 detainees being held at the Centre.. The identification requirements for visitors are very particular. Acceptable forms of Photo ID must be provided or two pieces of acceptable corroborating identification.

Different visiting conditions apply to two different categories of visitors, those classified by CIC as 'professional' visitors and as unprofessional visitors. No direct physical contact is allowed between detainees and unprofessional visitors. They must meet in the designated visiting area of Celebrity. They are separated by plexiglass and communicate through telephones. In addition to family and friends, visitors of the 'unprofessional' variety include:

Immigration consultants, church groups, doctors. Amnesty International personnel and other similar groups are not to be treated as Professional Visitors without prior approval from the EDO. They are not to be allowed

61 Post Orders, s.7.01
access into the Centre unless previous authorization has been obtained. These groups will use the detainee visiting area to conduct their business.

In contrast, 'professional' visitors may enter the inner regions of the facility. They include:

a) Lawyers
b) Embassy and Consulate Officials
c) Police
d) Interpreters
e) Approved Non-Government Officials, i.e. TRAC and Legal Aid with photo identification, the UN Commission for Refugees

Professional visitors are allowed into the Centre anytime to meet with detainees in the 'boardrooms' on the first floor.

In February, 2000, a new edition of the Celebrity's Post Orders was soon to be released. The draft of this new edition contains quite a few changes. One of them will amend the 'professional' and 'unprofessional' visitor categories. This is not surprising given that restrictions on access to the facility by Amnesty International 'and other similar groups' as well as by physicians has been particularly contentious. During the 1997-98 Citizenship and Immigration Standing Committee Hearings on Detention, critics charged that a letter was posted at the visit guard's desk which 'banned' Amnesty and other NGO groups. CIC officials denied the existence of such a letter: "For the record, there is no such list. Despite our best efforts, we're unable to find what this possibly was referring to."

In fact, what critics were referring to was not a 'letter', but rather the above quoted sections from the 1996 Post Orders, which were indeed posted at the Visits Guard's desk for at least as long as I was working there. It did not specifically 'ban'

---

62 Post Orders s.7.08

63 SIO interview. February 3, 2000

64 Neil Cochrane. Director, Case Presentation, Enforcement Branch, CIC. Citizenship and Immigration Standing Committee hearings, April 22, 1998:6
named ‘unprofessional’ groups, but rather ‘restricted their access’ to the facility. Further, the EDO could use their discretion under special circumstances to facilitate their admission into the facility. CIC’s denial was thus technically accurate - there was no ‘letter’ and there was no ‘ban’. However, without further clarification, it was clearly misleading. In the new forthcoming Post Orders, Amnesty International representatives and clergy have been upgraded to the category of ‘professional’ visitors. Doctors, Psychiatrists and other medical specialists who were deemed ‘unprofessional’ visitors in the 1996 Post Orders will now be treated as somewhat of a hybrid category. They will be allowed into the facility, however only with a referral or recommendation from the facility’s on-site doctor.\textsuperscript{65} The restrictions on the access afforded to consultants and paralegals will remain in force.

As with most of the rules and regulations of Celebrity, the EDO has the discretion to make and approve exceptions. With respect to the restrictions on visits by unprofessional, exceptions to their restricted access are most frequently granted for ‘humanitarian’ reasons. For example, an EDO might allow a detainee scheduled for imminent removal to meet directly with loved ones in order to be able to say goodbye more intimately. This is done ‘from time to time’.\textsuperscript{66}

c) Baggage and Valuables Access Schedule

Detainees may have access to their personal belongings being held by the Detention Centre’s authorities daily during the hours 18:00-20:00 or as authorized by the supervisor.

d) Telephone Time Schedule

Detainees can have access to a public telephone between: 9-10am. 2-3pm. 6-7pm. and 9-10pm or otherwise while in the dining rooms ‘...but not while food is

\textsuperscript{65}SIO interview. February 21, 2000

\textsuperscript{66}Ibid.
being served. The telephones are a vital part of the detainees' existence. Telephones connect them with the 'outside'. Contact with and access to friends and family members, lawyers and advocates, ethnic and religious community groups, embassies and shelters are critical for detainees who are seeking release. The phones are a precious and much vied for resource: people's lives are hanging in the balance. The demand for phones far exceeds their availability and their use is carefully governed. Long distance calls must be made from the pay phones, local calls may be made on the free phones. As a general rule, calls should not exceed 10 minutes. If a detainee requests to use a phone outside of the designated telephone times, a special request may be made and will generally be granted. 'Operational requirements permitting'. Incoming telephone calls may not be directly received by detainees but are received by security who "should obtain the name and the phone number of the caller, and pass the message to the detainee as soon as practical."68

e) Health Services

A medical doctor and nursing staff are contracted by CIC to attend to the health needs of the detainees. A nurse can be seen every morning between 9:00am and noon, and a doctor on Monday and Thursday during these same hours. Specialized health services, for example psychiatric counselling, may be provided if recommended by the doctor. Detainees are required to hand over all medications upon their detention and forthwith the onsite doctor has the responsibility of dispensing all medicines ('patent' or prescription') as needed. However, in the absence of medical staff, security personnel takes over this responsibility. 69 According to TRAC representative Eusavio Garcia, it is not uncommon to have the threat of transfer to jail used in the context of the provision of medical care. For example, a sick Chinese man was told by one immigration officer.

67 Post Orders. s.10.03

68 Post Orders. s.10.01

69 Post Orders s.15.03
that if he refused to take his medication he would be transferred to jail. 5 detainees who
had begun a hunger strike in January, 2000 were actually transferred to jail.⁷⁰

Since 1996, after the death of Michael Akhimen, any request to see the on-site
medical staff will be accommodated.⁷¹ On their scheduled days medical staff first see
those on their own lists of cases. Once they have seen everybody they want to see, they
will see as many "new" patients as time permits.

Not surprisingly, increased attention was paid to the health care conditions of
detention at Celebrity after the death of Akhimen. The Coroner made several
recommendations which CIC subsequently implemented, including first aid training for
the security guards and routine examinations of detainees held longer than a week.
However concerns about the quality of care persist. While the physical health of
detainees is unquestionably a primary concern of CIC officials and while access to on-
site medical services for physical illnesses has been improved, the mental health of
detainees remains a serious concern.

As a matter of policy, ill or injured detainees are transferred to a medical facility.
Exceptions are made "in minor cases".⁷² Many of those in detention exhibit serious
depressive symptoms and suicide is an ever present concern. The preferred way of
dealing with "suicide risks" is to transfer them to jail. If a detainee is deemed to be a
"low-risk suicide threat case" they are confined to one of the isolation rooms under
"special watch": "the person's status...will be checked every 10 minutes on all shifts."⁷³

---

⁷⁰Interview with Eusavio Garcia, TRAC Caseworker at Celebrity. February 3, 2000

⁷¹Pre-existing regulations gave the nurse the power to refuse any request to see medical
staff for "alleged" complaints. (Post Orders Appendix ‘H’) They even required medical
staff to "report abusive or hypochondriac detainees to the EDO. The EDO will, as
required, provide appropriate counselling to such detainees" (Ibid...). This is no longer
sanctioned by CIC. All requests to see the on-site medical staff must be accommodated.
(Interview with SIO at Celebrity February 3, 2000)

⁷²Post Orders s.15.01

⁷³Post Orders s.15.02
Generally speaking, CIC officials at Celebrity would prefer not to have either sick or suicidal people in their facility. Transfers out of Celebrity, either to a hospital or to a community health care facility for the sick or to jail for the suicidal, are sought wherever possible. The justification for the transfer of suicidal detainees to jail is that jails are better set up to deal with such cases; they have special ‘suicide watch rooms’. However, there is little question that jail is the last place that a seriously depressed or suicidal person should be. This reality is particularly pronounced in the case of refugee claimants who may have experienced torture and other human rights violations while being detained by authorities in their country of origin.

There are no provisions on-site for assisting detainees with psychological problems. While the medical staff may refer those with psychological problems to a counsellor, the practice is quite rare. It is more common is to transfer such cases to jail. A case recounted in a recent report on detention at Celebrity sheds some light on the attitudes which inflect the provision of health care at Celebrity. In 1997, a young woman refugee claimant, one of Celebrity’s many ‘long term’ detainees (over 30 days), was increasingly depressed and was suffering from serious physical symptoms. In response to queries made by the on-site TRAC representative regarding her health, medication and the possible need for a psychiatric referral, the on-site nurse vehemently argued “why should ‘we’ pay for a psychiatrist when she should be deported anyways”.

v) Numbers and Composition of Detained Population at Celebrity

The Celebrity Inn in the Toronto suburb of Mississauga is one of the three principal immigration holding centres in Canada. The other two are located in Vancouver, British Columbia and in Laval, Quebec. The facilities differ considerably in terms of the numbers of people detained and the conditions of detention. Vancouver utilizes a

74 Interview with SIO at Celebrity, February 3, 2000
downtown pre-trial facility and retains a small number of motel rooms at the airport. Laval’s facility, officially designated the ‘Immigration Prevention Centre’, is located in a refurbished former prison. Generally speaking, there are between 75 and 100 people detained at Celebrity which has the capacity to detain 100 people "comfortably". In contrast, Vancouver and Montreal hold approximately 20 to 40 people. 77

a) Who is in detention at Celebrity

As previously mentioned, immigration ‘holding centres’ are for the detention of non-criminal non-citizens: people who have come to the attention of the authorities, who have violated or are suspected of violating the Immigration Act and who are judged initially by Immigration officers and subsequently by adjudicators, to represent a ‘flight risk’. As explained by one CIC policy official, the people in detention at Celebrity and the two other holding centres “...are not people who have committed a crime. They’re not a danger, but they don’t have the right to remain in Canada... So we want to make sure we’re able to remove them.” 78 In this sense, detention is for deportation. "...the intention [of detention] with the non-criminal is that detention be as short as possible. In most cases it really is to facilitate removal, so there is not a justification for a long detention." 79

People detained in each of the three national holding centres include: those who have been deemed inadmissible to Canada and who have been refused entry (for a variety of non-criminal reasons including, for example, those suspected of not being ‘genuine visitors’); those who have been found working in Canada without authorization; those who have overstayed temporary visas and those whose refugee claims have been rejected. Approximately half of those detained at Celebrity are detained at the ‘front-end’ of the

76 Interview with SIO at Celebrity, February 3, 2000

77 "Is This Canada" 1998: 13

78 Greg Fyffe, Assistant Deputy Minister, Policy and Program Development. CIC Standing Committee Hearings, April 22, 2000: 20

79 Ibid..
immigration process (at the airport or other port of entry), and the other half at the 'back-end' (from within Canada, usually after a routine encounter with the police or other local authority).  

Ontario is the only region which regularly detains people with active refugee claims even when there is no criminal history.  

On February 3, 2000, for example, 10 of the 84 detainees at Celebrity (12%) had active refugee claims. The detention of asylum seekers is a matter of considerable concern. Not surprisingly, in recognition of the generally accepted inappropriateness of detaining refugees, CIC officials are quick to emphasize the distinction between the (genuine and deserving) refugee and (the always and already suspect) refugee claimant. CIC, they stress, does not detain refugees, only refugee claimants and only very few. As put by Minister of Immigration, Lucienne Robillard, "We put them in detention only if we have a serious reason to do so and once there, their conditions - again these are not prisons - are quite acceptable."  

In justifying the detention of refugee claimants, CIC observes that the practice is rare and exceptional:  

Of the 100 people being detained because its believed they would be unlikely to appear for their removal or for a hearing, approximately 15 of these people have an outstanding refugee claim [out of approximately 30,000 refugee claimants in Canada in total]. So we can see that it is unusual for a refugee claimant to be detained, and most of those cases involving a refugee claimant who is detained have circumstances that are unusual.  

80 Interview with SIO at Celebrity, February 3, 2000  
81 Ibid., 14  
82 Neil Cochrane. Director Case Presentation, Enforcement Branch. CIC Standing Committee Hearings. April 22, 1998:5  
83 Lucienne Robillard. Minister of Immigration, CIC Standing Committee hearings. November 27, 1997:15  
84 Cochrane. CIC Standing Committee Hearings, April 22, 1998:5
While the numbers of refugee claimants in detention at Celebrity at any one point in time is usually between 10 and 15 people, the numbers over time are slightly more significant. For example, TRAC case files indicate that during the 7 month period between September 1996 and March 1997, of those who sought the assistance of TRAC at Celebrity, 31 were front-end refugee claimants and 26 were rejected refugee claimants.

Moreover, the fact that the number of refugee claimants detained are generally small, does not diminish the particularly profound and debilitating effect of detention on refugees. Not only are their efforts to find, retain and communicate with legal counsel for their refugee claim as well as their ability to prepare their cases effectively severely restricted by their confinement, the impact of detention on their mental health and well-being can be debilitating:

The detention of refugees who have experienced violent persecution, often including imprisonment and torture, creates a set of circumstances which pose great risk to the well-being of the detained individual. The consequences of being detained again can cause traumatic affectations generally described as post-traumatic stress syndrome...a vulnerability to retraumatization exists....These psychological stresses severely limit the efforts of a person to establish a refugee claim. There is no recognition of their special circumstances and no medical treatment to meet their special needs.86

The vast majority of those detained at Celebrity are non-white, male, and between the ages of 20 and 40 years old. In early 2000, it was reported that China, Somalia, Iran, India and "the Caribbean" were commonly represented among the detainees.87

While statistics on the composition of the detainee population at Celebrity over time are not available, the following numbers provide a snapshot of the composition of the detainee population at Celebrity on one day, February 3, 2000. On this day there were a total of 85 people in detention. 32 (37%) were from Asia (China (14), India (7), Korea

85TRAC. "Is This Canada" 1998: 23
86TRAC. "Is This Canada" 1998: 23
87Interview with SI0 at Celebrity, February 3, 2000
were from Central and South America (Suriname (1), Guyana (4), Costa Rica (2), Venezuela (2), Mexico (4), Honduras (2), Argentina (1), El Salvador (2), Brazil (1). 14 (16%) were from Africa (Nigeria (5), Zimbabwe (3), Ivory Coast (1) Senegal (1), Somalia (1), Uganda (2), Namibia (1). 9 (11%) were from the West Indies (Dominican (2), St. Lucia (2), Jamaica (4), Grenada (1)). 8 (9%) were from Eastern Europe (Ukraine (1), Russia (2), Poland (1), Albania (1), Hungary (2), Uzbekistan (1). 1 person was from the Middle East (Iraq). 1 person’s national origin was unavailable. and 1 person came from western Europe (Netherlands). There were at least 5 pre-school age children in the detention centre as ‘guests’ of CIC, and at least one unaccompanied minor. The vast majority of those in detention were without resources, financial or otherwise.

While these numbers reflect the specific composition of the detainee population at Celebrity on one particular day, the general characteristics of the detainee population at Celebrity remain relatively stable over time. That is to say, the vast majority is always non-white, generally poor, mostly male. There are usually a handful of children with their mothers, one or two unaccompanied minors, one or two families, and about a dozen refugee claimants. While the specific countries represented at any one time vary, reflecting the shifting foci of CIC and police enforcement practices (the targeting of certain airlines, selective policing practices), the continental regions represented by the detainees remains relatively stable over time: Asia, Central and South America, Africa and Eastern Europe.

There is also some consistency over time with respect to the backgrounds of those detained. For example there are frequently eastern European sex trade workers at Celebrity, while many of the young women who come from the West Indies bear the scars of domestic violence. Many of the refugee claimants have fled the war-torn regions

---

88 CIC Detention and Removals Unit, Month End Statistics Report, 03-02-2000
89 Interview with Eusavio Garcia, TRAC Caseworker at Celebrity, February 3, 2000
90 Interview with SIO at Celebrity, February 3, 2000
of Eastern Europe and Africa, the repressive Chinese regime and Central and South American countries.

b) Length of Detention at Celebrity

CIC defines a ‘long-term’ detention as one which exceeds 30 days in length.\(^{91}\)

While most of the critical attention paid to long-term detentions focuses on those being held in correctional institutions for reasons of criminality, the concern is also present at Celebrity. While the Celebrity detention centre is designated for the short term, temporary detention of non-criminal non-citizens, many of those detained here remain for periods far exceeding 30 days.\(^{92}\) On February 3, 2000, for example, 40% (34) of the current detainees were, by CIC’s definition, long term detention cases. Of these 34 detentions, 11 had been detained for a period of 1-2 months; 6 had been detained for 2-3 months; 1 had been detained for 3-4 months; 8 had been detained for 4-5 months; 2 had been detained for 5-6 months; 5 had been detained for 6-7 months; and 1 had been in detention for 9 months. Of the 6 held in detention for more than 6 months, 3 were from China and 3 were from India.

Long term detentions at Celebrity often are related to identity and/or travel

\(^{91}\) Interview with SIO at Celebrity, February 3, 2000. This definition was also provided in the Citizenship and Standing Committee hearings, October 21, 1997:24 However, it should be noted that CIC’s definition of the official length of a ‘long-term detention’ is variable. CIC’s ‘Ontario Region Long Term Detention Statistics’ (31 December 1997) defines ‘long-term’ as ‘over 60 days’. And in the 1996-97 CIC Standing Committee Hearings, a ‘long-term detention’ was defined by Paul Thibault, Executive Director, Adjudication Division, CIC, as ‘over 90 days’. (March 31, 1998:3)

\(^{92}\) These concerns are shared by non-governmental critics as well as by the CIC officials at Celebrity who report that one of their major frustrations stems from the inappropriate fit of the actual physical facility (the Celebrity Budget Inn) and the nature and length of the detentions. Further, as reemphasized by a SIO at Celebrity, law and policy dictates that people should not be detained if there is no end to the detention in sight, either through release or removal. (See Chapter 2) However, despite this there are still many ‘long-term’ detentions at Celebrity.
documents. For example, adjudication often continues the lengthy detention of people who are slated for removal but who do not have the requisite identity/travel documents from their country of origin. These people often remain in detention while their applications are being processed by their national governments. Certain countries, for example India, are known for taking a very long time to produce the necessary documents. The longest term detainees at Celebrity are thus commonly citizens of India.

A long term detention at Celebrity may also be related to an active refugee claim. It is not uncommon for refugee claimants to be detained for the entire period during which their claim is processed and heard. While the IRB gives priority to detention cases, the wait can still be several months. As was the case for Michael Akhimen, this wait can be lethal. Like Akhimen, many refugee claimants are driven to withdraw their refugee claim even in the face of danger and loss of protection.93

The rationale for the continued detention of refugee claimants resembles a catch-22. TRAC has found that the more obvious it is to CIC officials and to adjudicators that the person seeking release from detention desperately wants to stay in Canada, the more likely it is that they will not be released because they are considered a flight risk. Refugees who honestly express their fear of returning home in their detention review are thus unwittingly strengthening the case for their continued detention. For this reason, TRAC routinely counsels detainees as to the difference between the detention review and the refugee hearing and emphasizes the importance of saving the details of their fears of persecution were they to be returned to their country of origin for their refugee hearing.

A long term detention may also be the result of a person refusing to sign the necessary papers in order to apply for the travel document. Indeed, CIC cites the "lack of detainee cooperation" in securing their travel documents, alongside the non-cooperation of the country of origin as being the primary reasons for long-term detentions, although they do admit that non-cooperation of detainees is a significantly smaller problem than

93TRAC. "Is This Canada" 1998:24
the non-cooperation of foreign governments. Nonetheless, one of CIC's primary preoccupations is to gain the 'cooperation' of detainees in the effort to apply for the necessary documents. Celebrity officials frequently use the promise of release from detention as an incentive to get detainees to 'sign'. CIC has also begun to attempt to facilitate this 'cooperation' during the course of detention reviews. As reported in the 1997-98 CIC Standing Committee hearings:

[CIC is]...currently working on initiatives to encourage our clients to comply with our requests for travel documents. For example, when we have clients who are detained, before they're released from detention our case presenting officers make arguments to the immigration adjudicator that the person should apply for a travel document before they are released from detention.

It is also the case that the threat of transfer to jail is sometimes employed as a means of coercing a detainee to 'sign'. In such cases, the refusal of a detainee to sign their application for a travel document is construed as evidence of non-cooperation with CIC officials and this 'unruliness' becomes the subject of a punitive response.

Many refugee claimants do not have the necessary identity/travel documents. For them, the consequences of this 'signing' has an additional edge. While the lack of requisite documents is commonly interpreted as an indication of a potential flight risk or possible danger and therefore justifies continued detention, the credibility of refugee claimants who agree to sign is immediately cast in doubt; "They say. you help us get a passport for you or we will keep you in detention. Then they help get a passport, and then the department turns around and says, you're not a refugee, because you applied for a passport."

94 Paul Thibault. Executive Director, Adjudication Division. CIC. Standing Committee Hearings. March 31, 1998:3
95 Interview with SIO at Celebrity, February 3. 2000
96 CIC Standing Committee hearings, February 18, 1998:10
97 David Matas. CIC Standing Committee hearings, March 25, 1998:20
The central importance of the signing of the travel document to immigration penalty bears a strong resemblance to the importance of a confession to sovereign penalty. Both legitimize and affirm the ‘rightness’ of the coercive policy and practices in question. Both facilitate and expedite these practices. Both render the subject of these practices complicit in their own subjectification. And, the withholding of both results in a more onerous challenge for the authorities who must work harder to achieve their desired objective. Moreover, the refusal to ‘sign’ is arguably the only bit of leverage available to detainees in their efforts to avoid deportation, for whatever reason. For all of these reasons, the ‘signing’ (in the context of immigration detention and deportation), like the ‘confession’ (in the context of criminal justice punishment), is a central and guiding preoccupation of authorities.

c) Release from Celebrity

Detention at Celebrity ends either with the release of the person into the community or the removal of the person from the country. As described in Chapter 2, an SIO may release someone from detention within the first 48 hours (the ‘48 hour review’). Once that 48 hours has passed, release may only be ordered by an independent adjudicator at the ‘detention review’. Usually release conditions are imposed usually including cash amounts or a performance bond, or both. Bonds have been set by adjudicators at levels ranging from 2,000 to 10,000 dollars. It is often the case that a person must remain in detention because they lack the resources to fulfill the release conditions. In this important respect, release from immigration detention differs from release from criminal incarceration. It is a fundamental and guiding principle of criminal justice bail reform that people should not be jailed for being poor. In the context of immigration detention, people are routinely refused release because they do not have enough money. Indeed, as observed by lawyer David Matas, the above mentioned principle of bail reform “...is totally absent from the Immigration Act. We have exorbitantly high bail amounts for immigration release and people do stay in immigration
detention because they are poor. That should not be so.\textsuperscript{98}

III) The Nature and Scope of Immigration Detention Outside of Celebrity

For the most part, Canada has resisted calls for a blanket increase in the use of immigration detention, choosing instead to target certain groups for increased detention (criminals and failed refugee claimants). In 1994, CIC detained just over 9,000 people (including 'criminals' and 'non-criminals').\textsuperscript{99} Since then the Canadian trend has been downward with respect to the total numbers of people detained. In 1996-97, CIC detained a total of 6,400 people for a total of 138,000 detention days.\textsuperscript{100} In 1997-98, CIC reported that they had detained just under 6,000 people in immigration detention, a decrease of 3,000 in 3 years.\textsuperscript{101} However, the length of individual detentions has steadily increased in Canada. For example in Ontario, the total number of immigration 'jail days' in provincial correctional institutions rose from just under 35,000 in 1992-93 to approximately 82,000 in 1995-96.\textsuperscript{102} Over the same period, the total number of 'holding centre days' in Ontario also rose dramatically. from 37,000 in 1992-93 to 48,000 in 1995-96. Since 1996, however, these numbers have been declining.\textsuperscript{103}

The 1995-96 increase in the lengths of jail and holding centre detentions reflects the heightened preoccupation with enforcement triggered by the 1994 Leimonis and Baylis shootings in Toronto. Significantly, the enforcement response which ensued ('criminals first') resulted in longer lengths of detention in both criminal and non-

\textsuperscript{98}Ibid..

\textsuperscript{99}Neil Cochrane, Immigration and Citizenship Standing Committee Hearings. submission. April 22, 1998:4

\textsuperscript{100}Standing Committee Hearings. April 22, 1998:20

\textsuperscript{101}Ibid..

\textsuperscript{102}Ontario Region Jail Day Histories, CIC Enforcement Statistics. 1997

\textsuperscript{103}Ibid..
criminal cases: in both correctional institutions and in holding centres.

The majority of immigration detention cases in Canada involve criminality issues. For example, in Ontario between 1994 and 1997, for every 1 person in detention at Celebrity, there were 3 being held in other correctional institutions.\(^{104}\) The majority of long-term immigration detentions are ‘danger to the public’ cases; cases in which the adjudicator judges that the release of the individual from detention would pose a danger to the public.\(^{105}\) As explained by CIC, “The dilemma the department is placed in is that if we release someone who is considered a high risk individual and they go out into public and kill a police officer or a citizen, we’re accountable for that.”\(^{106}\)

CIC reported in June 1997 that there were a total of 184 long term detentions in Ontario. 35 (20%) of these were non-criminal ‘flight risks’ being held at Celebrity. The remaining 149 (80%) involved criminality and danger issues and were being held in correctional institutions across the province. Of these 184 long term detentions, 83 (45%) had been detained for at least 6 months, 36 (20%) had been detained for at least a year and 9 (5%) had been detained for more than 2 years.\(^{107}\) The average length of detention for individual detainees before they were released or deported was 8 months (239.9 days).\(^{108}\) Of these 184 people, 86% (158) were non-white. 49% (90) were of African

---


\(^{105}\) These dangers to the public have been so found by an adjudicator. This finding does not necessarily mean that the Minister has declared them to be a danger to the public under s.70(5) of the Act.

\(^{106}\) Gerry Campbell. Assistant Deputy Minister. Operations. CIC. CIC Standing Committee hearings. October 21, 1997:24

\(^{107}\) CIC. ‘Ontario Region Long Term Detention Statistics’ 30 June 1997. The variable definition of ‘long-term’ detentions complicates the statistics. The numbers quoted here are based on the 30 day definition of a ‘long-term detention’. Curiously, the statistical report dated 31 December 1997, uses a 60 day definition of ‘long-term’ thereby making comparisons difficult.

\(^{108}\) Ibid.
descent. 39% (50) were Jamaicans and 20% (32) were Vietnamese.\textsuperscript{109}

On July 14, 1997 about 150 detainees at the Metro West Detention Centre, the facility where the majority of Ontario’s criminal immigration detainees are held, went on a hunger strike to protest the circumstances of their detention.\textsuperscript{110} Of particular concern was, and continues to be, the situation facing long-term immigration detainees. Typically, about 60% of the long-term detainees are permanent residents who have been declared by the Minister to be a danger to the public under s.70(5), and about 40% are failed refugee claimants who have been detained for reasons of criminality and who were judged by adjudicators to pose a danger to the public in their detention reviews.\textsuperscript{111} They have already served their criminal sentences, yet they remain in prison under ‘immigration hold’ while CIC attempts to effect their removal. Sometimes the time spent in prison in immigration hold exceeds the length of the criminal sentence served, a situation referred to as ‘double punishment’. For example, a refugee from Surinam who was granted refugee status in Canada in 1985, was convicted of several narcotics offences which netted him 4 months in jail. In July 1997, he had been on immigration hold for over 2 years.\textsuperscript{112}

At the time of the hunger strike, there were about 70 immigration detainees who had been detained at the Metro West Detention Centre in Toronto for more than 6

\textsuperscript{109} Ibid.. These statistics were submitted to the CIC Standing Committee on March 26, 1998

\textsuperscript{110}Parkdale Legal Services Inc., Memorandum re: “Metro West Detention Centre Hunger Strike by Immigration Detainees” July 22, 1997

\textsuperscript{111}Paul Thibault, Executive Director, Adjudication Division, CIC Standing Committee hearings. March 31, 1998:3 Thibault points out that the reasons for long-term detentions in criminal cases are similar for non-criminal cases: lack of detainee cooperation in securing the necessary travel documents or non-cooperation of the country of origin.

\textsuperscript{112}Ibid..
months. In addition to protesting against the ‘double punishment’ of long-term immigration detainees, the detainees were also protesting the “overcrowding and unsanitary conditions of their imprisonment: three inmates to a cell designed to sleep only two, so that one person has to sleep on the floor.”

In the United States, the INS has been busy building and opening new detention facilities. There, immigration detention is very big business. In 1986, The United States passed the Immigration Reform and Control Act. One of the consequences of this restrictive piece of legislation was the expanded use of detention for immigration violations. This trend was intensified with the passing of the 1996, Illegal Immigration Reform and Individual Responsibility Act. In 1995, the Internal Naturalization Service (INS) detained 6,000 people. In 1999, this figure had risen to 16,400. The average length of detention in the United States has also increased. In the United States, thousands of Mexicans are usually detained for a day or two before being deported. Still, hundreds of people have been held in INS detention for more than six months and dozens for more than a year. In 1994, the Americans spent approximately 200 million dollars (American) on immigration detention. Since the 1980s, it has opened more than 26 detention centres, some of which have been described as large ‘state-of-the-art’ facilities.

In Canada, government bureaucrats have developed a plan to build a similarly ‘state-of-the-art’ immigration detention facility which would be run by CIC and which would be equipped to ‘accommodate’ all varieties of immigration detainees: from failed

---

113 Parkdale Legal Services Memorandum, July 22, 1997
114 Ibid.
117 Ibid., 44
refugee claimants, to visa overstays to 'serious' criminal offenders.\textsuperscript{118} If this plan for the creation of a mega-detention complex is ever implemented, it would arguably represent the institutional manifestation of the culmination of the discursive developments which this thesis has traced: CIC would then be in the business of running 'prisons'. However, whether or not the government goes ahead with this plan remains to be seen. In the meantime, CIC continues to detain non-citizens in provincial jails and detention centres and in 'immigration holding centres.

The cost of detention is an extremely significant factor in the development of policy and practice in this area. While increasingly intensive enforcement practices imply more detentions, fiscal imperatives increasingly dictate restraint. In 1994-95, the Canadian government spent 21.1 million dollars on Immigration detention. In 1995-96, 23.4 million was spent on detention. In 1997, the national budget for detention was capped at 19.8 million.\textsuperscript{119} Approximately 13 million dollars is budgeted for immigration detention in Ontario. Of that figure, 4 million is designated for the 'Celebrity' detention centre which leaves 9 million dollars for the detention of criminals in Ontario.\textsuperscript{120} According to CIC, "...the cost to detain a person for a full year at the Celebrity Inn is $33,000.00. This is equal to $90.41 per day per person...Not included in this figure are the medical bills that range from 150,000 to 200,000 at the Celebrity Inn, or transportation costs (to immigration proceedings etc.)."\textsuperscript{121}

CIC observes that "[D]etention costs a lot of money...[that it] is used quite sparingly and with restraint...by the domestic regions across the country."\textsuperscript{122} And the numbers of immigration detainees are indeed relatively small; the numbers of non-

\textsuperscript{118}Interview with SIO at Celebrity February 3, 2000.

\textsuperscript{119}CIC Public Affairs. Response to Information Request. March 24, 1997

\textsuperscript{120}Ibid..

\textsuperscript{121}Ibid..

\textsuperscript{122}Gerry Campbell. Assistant Deputy Minister, Operations. CIC Standing Committee hearings. October 21, 1997:23
criminal detentions even smaller. So small in fact that the question of the real purpose and objective of immigration detention is necessarily raised.

This question becomes that much more pressing in light of the fiscal imperatives which ultimately govern detention and release decisions at Celebrity. Regardless of law, the bottom line is resources; “It is completely resource-driven. If there are no more beds, we pick up the phone and say don’t arrest any more people, don’t bring any more people here.”123 On one occasion, CIC temporarily gave up the use of 10 beds at Celebrity. When the beds became available again CIC officials were reluctant to get them back “...because they would just get filled up again.”124

In late February, 2000, there was no more room at the Inn. Celebrity was over capacity with 119 detainees. Arresting officers were told “...to stop arresting people we can’t remove, we can’t keep them here.”125 The resource-driven nature of detention decisions at Celebrity lead one SIO at Celebrity to conclude that detention at Celebrity serves no official policy purpose or objective whatsoever other than public relations: “...so Immigration can say ‘we’re doing something.’”126

While it may be true that immigration detention at Celebrity serves no official policy objective or purpose other than public relations, the importance of this exercise in public relations should not be underestimated. Immigration detention at Celebrity and elsewhere is a critical instrument for the maintenance and reproduction of coercive sovereign power and exclusionary sovereign objectives. In this analysis, detention and deportation are understood as ‘political rituals’ as ‘social phenomena’ which “cannot be accounted for by the juridical structure of society alone.”127 From this perspective the

123 Interview with SIO at Celebrity, February 3,2000

124 Ibid.

125 Ibid.

126 Interview with SIO at Celebrity, February 3,2000

127 Foucault Discipline and Punish p.24
relatively small numbers of people involved does not diminish the effects: while 'discipline' must act upon many people in order to be an effective governing power, sovereign power need only act upon a few, with the knowledge and assent of the many.

IV) Conclusion

Celebrity Inn is not a correctional facility, it is an Immigration 'holding' centre. The detainees are not convicted criminals, they are in fact specifically designated as 'non-dangerous' and non-criminal non-citizens. They are supervised and controlled not by prison guards, but by private security guards. Confinement at Celebrity is not carceral punishment, it is administrative detention. But this is indeed sovereign power in action. The forcible confinement of these individuals does not aim to 'correct', 'reform' or 'transform' souls. It has no official aim or purpose other than to confine, restrain and ultimately expel the actual bodies of non-citizens in accordance with the administrative exclusionary requirements of the Immigration Act.

Notwithstanding the longstanding legal distinction between administrative detention and deportation, and imprisonment and punishment, those who are suspected of violating the Immigration Act for whatever reason, criminal or otherwise, are constructed and acted upon as criminals (a trend which, as demonstrated in this thesis, has intensified over the last decade). And violations of the Immigration Act are acted upon through the application of a distinctly sovereign penalty, traditionally reserved for those convicted of criminal offences. The penalty to which Immigration violators are subject is one in which the 'body' (not the mind, or soul, or habits) is still the primary object of the sanction imposed, and the sanction is occasioned and justified by virtue of the juridical transgression of longstanding and resilient territorial sovereign law. In the context of immigration, detention and deportation are the most extreme, coercive and bodily sanctions available to sovereign states; sanctions which remain largely protected from serious judicial and/or public oversight precisely by virtue of this rationale. Needless to say, that these measures are legally defined as something other than punishment and that those being governed are legally acknowledged to be 'non-criminals' is largely irrelevant
to those who experience it as punishment and who are constructed and acted upon as if they were criminals, but who, ironically, are accorded fewer legal protections, fresh air, and support services than would be available to them if they had committed a crime.

Detention and deportation thus work on two distinct levels: as sovereign sanctions which work ritualistically and symbolically to produce and reproduce both the supreme and ultimately coercive power of the sovereign and of sovereignty and of the ‘rightness’ of the specific territorial sovereign policies and practices in question; and as sovereign sanctions which act upon, in coercive and discriminatory ways, the bodies of those constructed in the wider social and political context as undeserving and undesirable threats.

Sovereign penalty used to be a very public affair. Indeed, the effectiveness of punitive classical sovereign power depended upon the public ‘spectacle’ and ‘ceremony’ of bodily criminal punishment; sovereign penalty depends upon hegemony. However, physical pain and torture are no longer legitimate and hegemonic objectives of state sanctioned criminal punishment, and the spectacle of sovereign penalty has accordingly largely disappeared from criminal justice administration in Canada. As observed by Foucault, only ‘traces’ of sovereign power persist in the context of criminal justice, and these traces provoke shame and reform. Indeed, immigration detention and deportation is arguably the last remaining site where sovereign penalty is central and relatively unapologetic, largely unaffected by the humanitarian liberal influences on and reforms to the criminal justice system. In the context of contemporary exclusionary immigration, it is sovereign penalty that surrounds recently imposed ‘traces’ of liberal legality and humanitarianism.

In the late twentieth century, bodily sovereign power is no longer a public and shamelessly painful affair. Neither detainees and deportees nor convicted criminals are pilloried and paraded; notions of individual and human rights have dulled the sharp edge of the sovereign’s sword. However, while it is certainly true that in the context of Canadian Criminal Justice the body has been displaced as the object of punishment, in the context of exclusionary immigration law, policy and practice, the body continues to be
the constituent feature and object of the sanction. Immigration penalty entails, at its extreme, a fundamentally and inherently ‘bodily’ sanction: deportation. removal. banishment. exile: the physical subjection and expulsion of the offending ‘bodies’ in the name of the sovereign. There is no need to normalize or correct undeserving and/or undesirable non-citizens, through the operation of law, discretion, and sovereign power. they may be ceremoniously and force-fully expelled.
Epilogue: The Future of Immigration Detention in Canada

“Before the Law”

by
Franz Kafka

Before the law stands a gatekeeper. To this gatekeeper comes a man from the countryside and requests admittance to the Law. But the gatekeeper says that he cannot admit him just now. The man reflects and then asks whether he will be able to enter later on. “It is possible,” says the gatekeeper. “but not now.”

Since the gate to the law is open as usual and the gatekeeper steps aside, the man bends over to look through the gateway, into the interior. When the gatekeeper notices this, he laughs and says, “If it entices you that strongly, then try to go on in despite my prohibitions. But mind you: I am powerful. And I am only the lowest of the gatekeepers. And from hall to hall gatekeepers stand, each more powerful than the last. I myself cannot endure the sight of even the third one.

The man from the country did not expect such difficulties; after all, the Law, he thinks is supposed to be accessible to everyone. But when he now takes a closer look at the gatekeeper in his fur coat, his large pointed nose, the long, thin, black Tartar beard, he decides he would rather wait until he receives permission to enter. The gatekeeper Hands him a stool and lets him sit down at the side of the door.

There he sits for days and years. He makes many attempts to be admitted and tires the gatekeeper with his requests... Eventually his eyesight starts dimming, and he does not know whether his surroundings are actually darkening or his eyes are merely deceiving him. However, in the darkness he does recognize a glow breaking out inextinguishably from the door of the Law.

Before his death all his experiences of the entire period gather in his mind as one question that he has never asked the gatekeeper. He beckons to him since he can no longer raise his rigidifying body. The gatekeeper bends way down to him, for their difference in size has changed greatly to the man’s advantage.

“What else do you want to know?” the gatekeeper asks. “You are insatiable.” “All people strive for the Law,” says the man. “How come in these many years no one but me has asked to be let in?” “No one else could be let in here, for this entrance was meant for you alone. Now I’m going to go and shut it.”

In February 2000, the Enforcement Detention Officer at Celebrity was adamant. With 120 detainees on its register, there was absolutely no more room at the Inn. At these times, a phone call is made and CIC enforcement officers are directed to stop arresting people and delivering them to Celebrity. Rather than representing a meaningful and substantive check on the initial decision to detain, the '48 hour review' of the detention of new detainees (which is carried out on site by CIC officials at Celebrity within the first 48 hours of the initial detention decision, and which affords them the discretion to overturn the initial detention decision and to set release conditions) is used here as an important mechanism for the day to day control and management of the numbers of the detainee population at Celebrity. However, after this 48 hour period is over, the discretion of Celebrity officials to release is all but removed. Release decisions then become the prerogative of adjudicators whose mandate and concerns are different from those relating to the administration and management of the detention facility. Enforcement detention officers may then try to influence adjudication's decision, but they have no further direct control over release. So, when the beds are full, and there are no more 48 hour releases to be made, arrests and detentions are temporarily halted. The effective bottom-line at Celebrity with respect to detention and release decision-making is set by administrative and fiscal concerns; beds and budgets are the ultimate governing rationale here, not public policy objectives.

There are also the numbers to consider. At any one time, there are approximately 200 people detained in Canada by CIC for non-criminal reasons; most are at Celebrity (100-120 people; Vancouver and Laval usually house 20 - 40 people). While Celebrity is one of three locations across Canada where people are detained for non-criminal reasons, it is Canada's only semi-permanent 'Kafka hotel'. In terms of the size of its detainee population, it is the largest of the three facilities.

200 people. Even if one considers the frequent recomposition of this population through releases, deportations and new detentions, the annual detention figures of non-criminal are still small: criminal detentions outnumber non-criminal ones by
Of the 200 or so people in detention across Canada for non-criminal reasons at any one time, approximately 10 to 12 are likely to be refugee claimants detained at Celebrity (Ontario is the only region that regularly detains refugee claimants for non-criminal reasons). This represents a minute percentage of the approximately 30,000 refugee claims that are made on a yearly basis from within Canada.

So who are these people? It would be reasonable to assume that they are somehow exceptionally deserving of detention. After all, the decision was taken to detain them when the vast majority of similarly situated people are not, and indeed couldn’t possibly be detained. These 200 people, who are innocent of any criminal wrongdoing, are subjected under the authority of the Canadian Immigration Act to what is the most onerous form of punishment available under the criminal justice system: with, as we have seen, significantly fewer rights, protections and services than those accorded to citizens convicted of crimes. However, there is nothing particularly exceptional about these ‘prisoners’, except of course that they have committed no crime. Like the majority of Canada’s new immigrants and refugees they are non-white, and like the majority of people in prison for criminal offences, they are predominantly male and poor.

In February 2000, approximately 80% of those in detention at Celebrity were ‘long-term’ detainees (more than 30 days), a situation deemed most undesirable by the facility’s critics and administrators alike. Long-term detainees are brought before immigration adjudicators every 30 days for a review of their continued detention. While they have the legal right to representation at these reviews, in practice many attend them without any representation at all. Because of the reverse onus at the detention reviews, release is a difficult case to argue successfully, even with the assistance of legal counsel. It is also commonly known that, with or without legal representation, the chances of release by adjudication drop dramatically after a detainee’s first 30 day review as the rightness of the initial detention decision becomes entrenched over time. However, after a certain period, the chances of release for long-term detainees rise again. People who have been in detention for exceptionally long periods of time (usually for reasons relating to
bureaucratic delay and the lack of cooperation of foreign governments in issuing new identity documents required for deportation) are likely to be eventually released as it becomes clear that their case will not be resolved ‘imminently’ and therefore that their continued detention cannot be legally supported. As observed by an immigration official at Celebrity, many long-term detentions could and should be avoided by an early recognition that the resolution of the case at hand is likely to take some time due to well-known 'impediments' to removal, rather than waiting to make this decision until after a lengthy period of detention has been served.

The detention and release of non-criminal non-citizens at Celebrity is driven ultimately by fiscal and administrative imperatives, rather than by public policy objectives. Those who administer detention at Celebrity readily admit, and with more than a hint of cynicism, that there is no real policy objective being effectively served by the detention of these people at Celebrity and that the only real purpose of Celebrity detentions is public relations: "...so we can point at Celebrity and say we are doing something."²

In light of the above, it would make sense to close the Celebrity Inn immigration detention centre and stop the detention of all non-criminal non-citizens and refugee claimants across Canada. As criminal justice initiatives have shown, there are many non-carceral techniques for keeping tabs on people: indeed, in the face of budget restraints CIC has already been urging the expanded use of alternatives to detention for non-dangerous non-citizens. The evidence collected for this thesis would support a policy decision that CIC ‘holding centres’ be replaced by the use of non-carceral alternatives.

This sort of policy move would likely have more effect on the well-being of non-citizens who are not allowed to stay in Canada than potential legal moves to increase ‘rights’ and curtail administrative detention. As argued in this thesis, the liberal legal paradigm works to deflect critical attention away from unjust practices which, when they are addressed, are redefined and acted upon as legal problems requiring legal solutions.

²Interview with Enforcement Detention Officer at Celebrity. February 3. 2000
Consequently, much time, energy and resources is spent on resisting official policy and practices through the deployment of abstract legal arguments in formal legal and political contexts. Critical attention and resources are thus deflected away from material conditions and practices.

Moreover, there is a huge gap between the abstract formal existence of legal rights and protections for non-citizens in Canada, and the degree to which these rights may be accessed and operationalised to the concrete and material benefit of those who ‘bear’ them. This is not to say that liberal legal principles are not powerful and important potential checks on abuses, nor that they cannot be deployed in the resistance of unjust practices, for certainly both are true. However, the effective use of legal rights and protections in the pursuit of justice depends upon the ability of the ‘rights bearer’ to access and deploy them. Access is determined largely by resources. Non-citizens in general lack resources and issues for non-citizens in detention are obviously magnified. They generally lack the economic, political and legal supports, resources and contacts which are necessary in order to effectively mobilize legal discourses in the resistance of forms of subjectification. Without effective and practical access to liberal rights and protections, their abstract existence is at best irrelevant to those in detention, and at worse, provides ideological legitimation for coercive and unjust practices while simultaneously deflecting critical attention away from them.

This study is, however, somewhat at odds with current CIC thinking and indeed with international trends. In Canada, there is now a proposal to build or buy a new ‘mega-detention centre’ in Ontario in which CIC detainees would be confined. Those who are currently serving time in prison for criminal offences would be transferred to immigration hold at the new detention centre upon the completion of their sentences. The new centre would accommodate both ‘dangerous’ and ‘non-dangerous’ non-citizens in medium and maximum security facilities under one roof. It would be large, modern and equipped with state of the art surveillance technology. It would be specifically designed to meet all national and international formal legal and human rights requirements regarding detention conditions. Immigration detention would be stream-lined, centralized and standardized. In
short. CIC would enter the business of administering prisons.

Canada is not alone in this preoccupation. Since the late 1980s, there has been a general increase in immigration detention and in the detention of asylum seekers in the United States and in most western European countries.3 Western Europe has also seen the proliferation of increasingly restrictive immigration legislation, including, for example, the 1993 ‘Aliens Act’ in Germany, the 1995 ‘Federal Law on Security Measures to be Taken in Relation to Foreigners’ in Switzerland, the 1994 French ‘Waiting Zone’ detention legislation, the 1996 Asylum and Immigration Act in the United Kingdom and the 1995 Danish ‘Aliens Act’. Much of the legislation has expanded the scope for immigration detention and has therefore been accompanied by plans in a number of countries to open new detention centres or to extend existing ones especially for this purpose. For example, in the United Kingdom, a new, ‘purpose-built’ detention centre was opened in May 1996 at Gatwick Airport4, and the passage of a proposed 1999 Immigration and Asylum Bill would establish a system of ‘cashless support’ for asylum seekers in ‘designated’ accommodation around the country.5 As reported by Amnesty International, since the opening of new detention facilities in the United Kingdom, the number of immigrant and refugee detentions has steadily increased. Similar trends are evident in France, Germany, Belgium and Switzerland.6

While the building of a mega-detention centre in Ontario might be an improvement for those currently languishing in the Don Jail or other correctional facilities under ‘immigration hold’, and while there is little question that the current conditions of immigration detention are less than adequate, we know from the literature

3Inter-Church Committee for Refugees, “Detention in Europe and Legal Aspects of Challenges to Detention in the United Kingdom” Workshop on Detention. July 7, 1997 (mimeo)

4Ibid..


6Ibid..
and experiences of criminal justice that the building of new, large and modern carceral facilities results in an increase in imprisonment, a concern as mentioned above, already expressed in the context of immigration detention in the United Kingdom. Relatedly, this policy move would also serve to legitimize immigration detention and deflect critical attention away from it, in a way that is not possible under current conditions.

Whatever the eventual outcome of CIC’s current policy deliberations, efforts should be made to ensure humane conditions and practices prevail in the future governance of the detention of non-citizens. Short of an end to the use of detention for non-criminal non-citizens, several features are critically important in this regard and should at the very least form a part of any new policies and practices relating to immigration detention in Canada. These include:

a) The creation of an external independent civilian review process for all CIC enforcement activities;

b) The creation of an independent complaints mechanism for people in detention;

c) Regular inspections of immigration detention facilities by an independent human rights authority;

d) Meaningful and substantive access to legal counsel, interpreters, information, services and supports;

e) Fresh air and exercise;

f) Direct contact between detainees and their visitors;

g) The provision of independent, culturally-sensitive, medical services (including comprehensive provisions for identifying and responding to the physical and psychological needs of people in detention and an end to the practice of transferring ‘suicide risks’ to jail);

h) The end to the indiscriminate use of body restraints on detainees who are being transported or who are appearing before a court or tribunal.

i) The provision of intensive, cross-cultural, anti-discrimination and human rights training for all those engaged in the enforcement functions of CIC, including immigration officers, private security officers and medical personnel.
Sources of Information

Interviews


Enforcement Detention Officer at the ‘Celebrity Inn’ Immigration Holding Centre in Mississauga. Toronto: February 3, 2000; February 20, 2000


Purvis. Rob. Constable, Morality Squad Unit, York Region Municipal Police Department. Toronto: August 20, 1999

Redmonds. Maggie. Former Community Development Officer for Metro Social Services. Toronto: May 6, 1999

Conversations

In addition to the above listed formal interviews, and of course to conversations with the teaching staff of the Centre of Criminology at the University of Toronto (including in particular the members of my thesis committee) and with fellow students at the Centre, conversations with several individuals were particularly helpful in instrumental and insightful ways; Fred Franklin, Detention Committee, Toronto Refugee Affairs Council (TRAC); Ursula Franklin, Professor Emeritus, University of Toronto; Nicola Lacey, Fellow and Tutor in Law, New College, Oxford; Peter Fitzpatrick, Professor of Law and Social Theory, University of Kent.

Conferences, Meetings, Hearings Attended

Citizenship and Immigration Standing Committee Hearings on Detention and Removals.

Inter-American Commission on Human Rights, Canadian Deputations on Detention.
Toronto, October 22, 1997.


NGO-CIC Working Group Meetings, Toronto: October 14, 1997; December 12, 1997; and March 11, 1998.

Toronto Refugee Affairs Council (TRAC) meetings with Immigration Enforcement officials responsible for detention at the 'Celebrity Inn' Immigration Holding Centre in Mississauga, Ontario: August 20, 1997; November 20, 1997; and January 6, 1998.
Field Work

Between August 1997 and June 1998 I volunteered as the on-site case worker for the Toronto Refugee Affairs Council (TRAC) at the ‘Celebrity Inn’ Immigration Holding Centre in Mississauga. TRAC has an agreement with CIC which provides office space and telephone access for a case worker who provides information and referral services to detainees. Such case workers are specifically prohibited from performing case-specific advocacy services: they are there to ‘facilitate access to the system’. I performed this function for TRAC, two days a week for 9 months. I am particularly grateful to Fred Franklin of TRAC for having facilitated and supported my work at Celebrity, and to the CIC officials at Celebrity who for the most part responded to my queries and tolerated my curiosity while I was there, and who were open and forthcoming in subsequent interviews. While I was not able, for ethical reasons, to use any of the case-specific information that detainees provided to me during the course of many interviews, observations were nonetheless plentiful.

Print and Broadcast Media

I collected newspaper clippings on immigration and refugees, criminality and exclusions and welfare issues from September, 1993 and December 1998. I draw largely, but not exclusively, from The Toronto Sun and The Globe and Mail. I also used the transcripts of several CBC Radio Broadcast of relevant news stories including: the use of profiles in the granting of overseas visa applications (September 9, 1998); the deportation of a long-term permanent resident deemed to be a ‘danger to the public’ (1997); the forging of official deportation papers by immigration officials.
Works Cited

Cases

*Azam.* [1973] (2 All E.R. 741)

*Chiarelli v. Can. (Minister of Employment and Immigration).* 16 Imm.L.R. (2d) 1, [1992] 1 S.C.R. 711


*Ribic v. Canada (Minister of Employment and Immigration)* (unreported, August 20, 1985. I.A.B., docket no. 84-9623)


*Williams v. Minister of Citizenship and Immigration* April 11, 1997 A-855-96 (IMM-3320-95)

Statutes

*Immigration Act.* 1976-77, c.52, s.1

*Canadian Charter of Rights and Freedoms.* 1982

*Canadian Security Intelligence Service Act.* 1984

An Act to amend the Immigration Act and other Acts in Consequences Thereof. (Bill C-86). 1992

An Act to Amend the Immigration Act and the Citizenship Act and to make a Consequential Amendment to the Customs Act. (Bill C-44). 1995.

The Controlled Drug and Substances Act, (Bill C-8) 1996

Canadian Government Publications and other Documents

Canadian Security Intelligence Service. “The CSIS Mandate” CSIS Website:

------. “Operational Programs: Security Screening” CSIS Website:


------. “Submission to the Special Committee of the Senate on Security and Intelligence” by Ward Elcock, Director of CSIS, June 24, 1998 (mimeo).

------. 1996 Annual Report on Organized Crime in Canada


------. Detention and Removals, Mississauga Immigration Holding Centre Post Orders. February 29, 1996.


------. Citizenship and Immigration Detention Policy (DRAFT) 1996


CIC Update. Ottawa, No.8, April 1997.

Enforcement Branch. Ontario Region "Enforcement Statistics on Detention and Removals". 1997 (mimeo)

Enforcement Branch, Operations Memorandum. EC- 98-08. October 27. 1998

Immigration Holding Centre, Mississauga, Rules for Detainees. 1998

Public Affairs. Ontario Region, Information for Persons Detained. 1998


Detention and Removals Unit, Month End Statistics Report. February 3, 2000


White Paper on Immigration, Ottawa: October 1966

Managing Immigration: A Framework for the 1990s, Ottawa: Supply and Services Canada, 1992:

Explaining the New Immigration Bill, Ottawa, 1992


Solicitor General of Canada. “Notes for an Address by the Honourable Doug Lewis. P.C., M.P., Minister of Public Security, , August 9, 1993

*International Documents*


------. *UNHCR Guidelines on Detention of Asylum Seekers* PRL 5.3. 29 January 1996


Articles and Books


Amnesty International. "Brief to the Parliamentary Standing Committee on Citizenship and Immigration" (mimeo), 1998

---. Annual Human Rights Report 1997


Auerbach. P. Harris. "Discretion and Section 19(1)(a) of the Immigration Act" Journal of Law and Social Policy, 6, 1990:133-147

------. 'Dangerous Foreigners': *European Immigrants Workers and Labour Radicalism in Canada. 1896-1932* Toronto: McClelland and Stewart, 1979


Canada Employment and Immigration Union. Public Service Alliance of Canada. “Presentation to the House of Commons Standing Committee on Citizenship and Immigration”. March 25, 1998 (mimeo)

Canadian Bar Association, National Immigration Law Section. Submission on Bill. C-44. November 1994


-------. “Refugee Detention in Canada” (mimeo), 1994


-------. “CCR Calls on Canadians to Respond Warmly to Refugees” May 6. 1999 (mimeo)


Dworkin. Ronald. “Judicial Discretion” *60 J. Philosophy* 1963

--------. *Taking Rights Seriously* Cambridge, Mass.: Harvard University Press. 1977


Falconer, Julian and Carmen Ellis "Colour Profiling: The Ultimate Just Desserts" (draft) 1998 (mimeo)


---------. *Immigration Law in Canada*, Toronto: Butterworth and Co., 1984


Hannah-Moffat, Kelly. "Moral Agent or Actuarial Subject: Risk and Canadian Women’s Imprisonment” (Draft) 1998


Inter-Church Committee for Refugees “Detention in Europe and Legal Aspects of Challenges to Detention in the United Kingdom” Workshop on Detention (mimeo). July 7, 1997


Leddy, Mary Jo “Comments on Detentions and Removals: Brief to the Standing Committee on Citizenship and Removals”, 1998 (mimeo).


Malarek, Victor. Haven’s Gate: Canada’s Immigration Fiasco. Toronto: Macmillan. 1987


Matas, David “Immigration Removals: A Submission to the House of Commons Standing Committee on Citizenship and Immigration”. March. 1998 (mimeo)

Metro Toronto Chinese and Southeast Asian Legal Clinic. “Submission to the Standing Committee on Citizenship and Immigration. 1998” (mimeo)


------. *Turning Point*, Toronto: Government Printer, 1993


Pacheco "Ministerial Discretion Under S.37 of the Immigration Act. 1976"  

Parkdale Legal Services. Memorandums on Immigration Detention and Deportation.  
Toronto: 1997


Riley. Diane. “Drugs and Drug Policy in Canada: A Brief Review and Commentary”  
Canadian Foundation for Drug Policy, November 1998 (mimeo)

-------. “Drugs and Drug Policy in Canada:A study prepared for Senator Pierre Claude Nolin as a background document for his June 1999 motion to have Canada’s Senate conduct a thorough review of Canadian drug law and policy”. Canadian Foundation for Drug Policy. 1999

Roberts, Barbara. *Whence They Came: Deportation From Canada 1900-1935*. Ottawa:  
University of Ottawa Press, 1988

Roberts. Julian and Anthony Doob. “Race, Ethnicity and Criminal Justice in Canada”.  

Rose. Nikolas “Government Authority and Expertise in Advanced Liberalism” in  
*Economy and Society* Vol. 22, No. 3. August 1993:283-299


------. Poor Discipline. Chicago: University of Chicago Press. 1993


Somali Immigrant Aid Association. Background Information. October 1999 (mimeo)


Stoffman, Daniel. "Dispatch from Dixon" Toronto Life August 1995

------. "Making Place for Real Refugees" International Journal Vol 52 (4) 1997:575-580

Strange, Carolyn (ed.) Qualities of Mercy: Justice, Punishment and Discretion Vancouver: University of British Columbia Press, 1996


Toronto Refugee Affairs Council. “Brief to the Government of Canada about Recommendations by the Legislative Review Committee” March 1998 (mimeo)


Valverde, Mariana “‘Despotism’ and Ethical Governance” Economy and Society 25(3) 1996:357-372

Valverde, Mariana et al (eds) Wife Assault and the Criminal Justice System. Toronto: Centre of Criminology, University of Toronto, 1995


Young, Alison. Imagining Crime London: Sage Publications. 1996
INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

Bell & Howell Information and Learning
300 North Zeeb Road, Ann Arbor, MI 48106-1346 USA
800-521-0600

UMI
EXPLORING AGE AND MATURITY IN YOUTH JUSTICE

Kimberly N. Varma

A thesis submitted in conformity with the requirements for the Degree of Doctor of Philosophy

Centre of Criminology, University of Toronto

© Copyright by Kimberly N. Varma (2000)
The author has granted a non-exclusive licence allowing the National Library of Canada to reproduce, loan, distribute or sell copies of this thesis in microform, paper or electronic formats.

The author retains ownership of the copyright in this thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without the author’s permission.
EXPLORING AGE AND MATURITY IN YOUTH JUSTICE

Kimberly N. Varma
Degree of Doctor of Philosophy
2000
Centre of Criminology, University of Toronto

ABSTRACT

This thesis examines the ways in which decisions made about youths in conflict with the law are based upon, or affected by, notions of their chronological age and apparent maturity. An empirical examination of youth court hearings, youth court data, and a survey of the Ontario public was undertaken in order to explore this question. The findings reveal that age and apparent maturity appear not to be related to decisions made about youthful offenders. However, ‘youthful’ status was recognized in the courts when it came to the level of supervision that was necessary to manage youths.

This study provides a view to understanding the social construction of adolescence and the underlying ambivalence toward youth who come in conflict with the law. The following account of decision-making about young offenders by court practitioners and members of the public reveals the arbitrary nature and fluidity of categories of age, and also allows us to explore how ‘youth’ plays out in the youth justice system.
ACKNOWLEDGEMENTS

In undertaking this project, I have been fortunate enough to come into contact with many individuals who have provided an enormous amount of support and encouragement. First, this research would not have been possible without the assistance of individuals at the various youth courts in Toronto. I owe a special debt of gratitude to Judge Lynn King, Jennifer Kerr, Judge Joseph James, the staff at 311 Jarvis street, John Dick, as well as Judge David Cole for their advice, guidance and assistance. I also would like to thank, in particular, Sutapa Karmakar for having the patience and energy to assist me in collecting these data, and doing so in such a careful and meticulous way. In addition, I am grateful to Goldfarb and Associates for the collection of the public perceptions data, as well as Bert Steenburgh, Marg Stanowski and the staff at Operation Springboard for their interest and support for this project. In addition, I would like to extend my appreciation for the support of this research made possible by a SSHRC grant to A.N. Doob at the University of Toronto.

I have also been extremely fortunate to have crossed paths with incredibly talented and helpful individuals at the Centre of Criminology. Rita Donelan, Monica Bristol and Gloria Cerin vivo have been exceptionally supportive. Thank-you for asking if I was eating right and getting enough sleep – I now realize the importance of these things. Tom Finlay and Andrea Shier were both extremely patient and helpful. Thank-you for renewing that one book I had for two years straight. I also could not have asked for a more spirited faculty at the Centre of Criminology. John Beattie, William Watson, Philip Stenning, Clifford Shearing, Scot Wortley, Carolyn Strange. I have learned such a great deal from all of you. I would also like to especially say thanks to Rosemary Gartner and Jim Phillips for being so incredibly supportive and encouraging. I am also very grateful to have worked with Dick Barnhorst. I am thankful for all of his advice and encouragement over the years.

I am grateful to have shared this experience with Renisa Mawani, Anna Pratt, Kirsten Johnson, Jennifer Wood, Cheryl Webster, Andrea MacKenzie, Simon Devereaux, Greg Smith, Kimberley White-Mair, Mary-Lynn Young, Phil Mun, Carla Cesaroni, Michael Kempa, and Dawn Moore. I have also been fortunate to have been supported by those that have come before me: Tammy Landau, Willem Delint, Benedikt Fischer and in particular Kelly Hannah-Moffat who has helped me in so many ways. To Jane Sprott and Voula Marinos, I owe my sanity (or insanity). Thank you for your friendship.

Thanks to my family and friends: Raj and Joe Varma, Jyoti and Art Zuidema, Dee Varma, Judy and Roger Barker, Lynda and David Hunter as well as Jennifer Newton and John Paul Cromien.

My committee has been an incredible source of inspiration. Thank-you to both Kent Roach and Mariana Valverde for your encouragement and support from the beginning to the end. As well, I extend my appreciation to Tullio Caputo for his comments and helpful advice.
I am deeply appreciative of having the opportunity to have worked with Tony Doob; an incredible supervisor, mentor and friend. I thank you for your patience, humour, and most of all your constant support. I could not have asked for a better supervisor...thankfully I asked you.

I will end by expressing my deep sense of gratitude to Ian Barker. You have been completely supportive and encouraging throughout this process. Little did you know back in 1988 that this is what you were getting into. Thank you for your patience, and understanding. I could not have done this without you.
# TABLE OF CONTENTS

| ABSTRACT | ii |
| ACKNOWLEDGEMENTS | iii |

## CHAPTER ONE
Introduction ........................................................................................................... 1
Present legal context .................................................................................................. 7
Historical Context and the Process of Reform .......................................................... 9
Developmental Research .......................................................................................... 18

### CHAPTER TWO - BAIL HEARINGS
Introduction .................................................................................................................. 23
Legal Context ................................................................................................................ 24
The Decision-Makers ................................................................................................... 28
Research Methods ....................................................................................................... 29
The Sample of Observed Cases .................................................................................... 31
Results: The Crown ..................................................................................................... 34

### Section I - Legal factors relating to the Crown decision

- Legal Index .............................................................................................................. 35
- Principal Charge .................................................................................................... 36
- Prior Record .......................................................................................................... 37
- Bail Abuse and Previous Pre-Trial Detention ......................................................... 40
- Reverse Onus ......................................................................................................... 40
- Co-Accused in the Present Offence ........................................................................ 43
- Discussion of legal factors ...................................................................................... 44

  - 'Child-like' vs. 'Adult-like' Index ........................................................................ 45
  - Age ........................................................................................................................ 46
  - Gender ................................................................................................................... 47
  - How the youth 'appeared in court' ....................................................................... 48
  - Adult or Child-like signs ....................................................................................... 49
  - Discussion of Demographic and 'Child-like' vs, 'Adult-like' factors ....................... 49

- Support Index ........................................................................................................... 50
- Family and School .................................................................................................... 51
- Living Arrangements ............................................................................................... 51
- School Attendance ................................................................................................. 53
- Discussion of Support Factors ................................................................................ 55

- Which factors overall predict the Crown decision on bail release? ....................... 55
- Conclusions – the decision making process of the Crown ...................................... 56

### Section II - Factors affecting the decision to grant bail on Contested cases
Legal Index .................................................................................................................. 57
Principal Charge ................................................................. 59
Prior Record ........................................................................ 60
History of Bail Abuse and Previous Pre-Trial Detention ........................................ 61
Reverse Onus ....................................................................... 62
Co-Accused in the Present Offence ............................................ 62

‘Child-like’ vs. ‘Adult-like’ Index .................................................. 63
Demographic Factors in the Bail Decision ........................................ 64
Other Demographic and Appearance related factors ............................... 65

Support Index ........................................................................ 66
Parental support ...................................................................... 67
Living Arrangements and School ............................................... 67
Discussion ............................................................................... 68

Section III – The Overall Detention Decision
Legal Index ............................................................................ 69
Principal Charge ..................................................................... 71
Prior Record ........................................................................... 72
Bail Abuse and Previous Pre-Trial Detention .................................. 73
Reverse Onus .......................................................................... 74
Co-Accused in the Present Offence ............................................. 74

‘Child-like’ vs. ‘Adult-like’ Index .................................................. 75
Demographic factors in the bail decision ........................................ 76
Age ..................................................................................... 76
Other Demographic and Appearance Related Factors ............................ 78
Discussion of Demographic and ‘Child-like’ vs. ‘Adult-like’ factors .......................... 78

Support Index ........................................................................ 78
Parental support ...................................................................... 79
Living Arrangements ................................................................ 79
School Attendance ................................................................... 80
Conclusions Section III .............................................................. 81

Section IV – Conditions of Release ............................................... 83
What conditions were generally placed on the youth in this sample? .................. 86
Conclusions ........................................................................... 87
How is the youth constructed and by whom? ......................................... 89

Appendix A – Bail Coding Sheet .................................................. 91
Appendix B – Inter-Rater Reliability ............................................... 96
Appendix C – Breakdown of Indices ................................................. 97
# Chapter Three - Sentencing Hearings

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Introduction</td>
<td>98</td>
</tr>
<tr>
<td>Legal Context</td>
<td>99</td>
</tr>
<tr>
<td>Research Methods</td>
<td>100</td>
</tr>
<tr>
<td>The Sample of Observed Cases</td>
<td>102</td>
</tr>
<tr>
<td>Section I - Results</td>
<td>104</td>
</tr>
<tr>
<td>Legal Index</td>
<td>104</td>
</tr>
<tr>
<td>Pre-trial Detention</td>
<td>105</td>
</tr>
<tr>
<td>Principal Charge</td>
<td>107</td>
</tr>
<tr>
<td>Prior Record</td>
<td>108</td>
</tr>
<tr>
<td>Co-accused in the case</td>
<td>112</td>
</tr>
<tr>
<td>Role of youth in offence</td>
<td>113</td>
</tr>
<tr>
<td>Child-like/Adult-like Index</td>
<td>114</td>
</tr>
<tr>
<td>Age Groups and Custody Across Canada</td>
<td>117</td>
</tr>
<tr>
<td>Breakdown of Demographic Factors - Ethnicity</td>
<td>125</td>
</tr>
<tr>
<td>Other Demographic and Appearance Related Factors</td>
<td>128</td>
</tr>
<tr>
<td>Support Index</td>
<td>131</td>
</tr>
<tr>
<td>Parents</td>
<td>132</td>
</tr>
<tr>
<td>Living Arrangements</td>
<td>133</td>
</tr>
<tr>
<td>School</td>
<td>134</td>
</tr>
<tr>
<td>Learning Disabilities</td>
<td>134</td>
</tr>
<tr>
<td>Employment</td>
<td>134</td>
</tr>
<tr>
<td>Legal Representation</td>
<td>135</td>
</tr>
<tr>
<td>Case Plans</td>
<td>136</td>
</tr>
<tr>
<td>Joint Submission</td>
<td>138</td>
</tr>
<tr>
<td>Participation of youth and parent(s) in sentencing hearings</td>
<td>139</td>
</tr>
<tr>
<td>Conclusions of Section I</td>
<td>140</td>
</tr>
<tr>
<td>Section II - Principles and Goals of Sentencing</td>
<td>141</td>
</tr>
<tr>
<td>Cases that Received Custody</td>
<td>144</td>
</tr>
<tr>
<td>Section III - Probation</td>
<td>146</td>
</tr>
<tr>
<td>Probation Conditions</td>
<td>147</td>
</tr>
<tr>
<td>Age Group</td>
<td>149</td>
</tr>
<tr>
<td>Number of Conditions Placed Upon Youth by Age</td>
<td>150</td>
</tr>
<tr>
<td>Section IV - Pre-Disposition Reports</td>
<td>151</td>
</tr>
<tr>
<td>Child-like/Adult-like Language used in PDR's</td>
<td>156</td>
</tr>
<tr>
<td>Conclusions</td>
<td>160</td>
</tr>
<tr>
<td>Appendix A - Sentencing Coding Sheet</td>
<td>162</td>
</tr>
<tr>
<td>Appendix B - Inter-Rater Reliability</td>
<td>167</td>
</tr>
</tbody>
</table>
Appendix C – Breakdown of Indices ................................................................. 168
Appendix D – Detailed information on cases where reference was made to
principles and/or goals of sentencing .............................................................. 169
Appendix E – Detailed information on references to S.24 of the YOA for
cases that were considered for custody ......................................................... 173
Appendix F – Detailed Analysis on Pre-Disposition Reports .......................... 174

CHAPTER FOUR – PUBLIC PERCEPTIONS

Introduction ......................................................................................................... 187
Research Methods .............................................................................................. 190
Section I – Examining the effects of information about young offenders ........ 190
Results – sentencing preferences ..................................................................... 192
Ratings of the characteristics of the youth ....................................................... 197
The interactions of ratings of the youth and sentencing preferences .............. 199

Section II – Does the public differentiate between youths and adults
on the purposes of sentencing? ......................................................................... 201
Comparing youths and adults – goals of sentencing ....................................... 203
Views of the purposes of sentencing by other beliefs
about the justice system .................................................................................... 205
Views of the purposes of sentencing by beliefs about
severity in sentencing ....................................................................................... 207
Conclusions ........................................................................................................ 209

CHAPTER FIVE - CONCLUSIONS

How do these models play out in the youth justice system? .......................... 213
Summary of Findings ......................................................................................... 214
The Disappearance of youth? ........................................................................... 215
Remembering that there are differences among youth .................................. 216
"Re-establishing" youth as a concept ................................................................. 217
Conclusion ......................................................................................................... 220

REFERENCES .................................................................................................... 221
Chapter One

Introduction

This thesis examines the ways in which decisions made about youths in conflict with the law are based upon, or affected by, notions of their chronological age and apparent maturity. To this end the following chapters will provide an empirical examination of various sites -- youth court hearings and a survey of the Ontario public -- in order to explore how constructs of age and maturity affect decisions made about youth who break the law. My hypothesis was that decisions made in court and by the public would be related to age. Specifically, I expected to see differences in the treatment of younger versus older youths. As will become apparent, I found little evidence to substantiate such a claim. As a result the present study seeks not only to understand the factors that are related to decisions made about youth who break the law, but also attempts to understand why age does not appear to relate to decisions made in youth court proceedings, nor to judgements made by the public.

While the purpose of this thesis is quite straightforward, the findings that will be presented have implications beyond a simple examination of the effects of age on decision-making in a criminal justice context. By examining if ‘youthfulness’, or a young person’s state of maturity, factor into decisions made about their criminal responsibility we are in a better position to evaluate the assumptions underlying the regulation of young people in other domains. Understanding decisions made about ‘youthfulness’ in this study provides insight into the social construction of adolescence more generally, and also provides an understanding into the broader topic of the
adjudication of different populations as morally autonomous. For instance, the way in which criminal responsibility is determined in insanity proceedings or in terms of the use of battered women syndrome as a legal defense sheds light on how notions of mitigated criminal responsibility are constructed and decided upon. While this thesis will not specifically examine the connections between youth in conflict with the law and these other groups, the findings may be of interest to researchers who are concerned with the ways in which mitigated criminal responsibility is defined in other contexts.

An interesting example of how adolescents are regulated differently depending upon the context comes from recent initiatives undertaken by the Federal Minister of Justice, Anne McLellan, to raise the age of sexual consent to sixteen from the age of fourteen where it has been for nearly 120 years.¹ This proposal for change is apparently a response to current concerns over the protection of young people under the age of sixteen from being victimized by adult sexual predators. At the same time, however, the proposed change will have legal consequences for young people under the age of sixteen who are consenting partners engaging in sexual activity. The assumption underlying this proposal is that fourteen and fifteen year olds may not be mature enough to understand decisions made about sexual relations. As stated in the Justice Department’s consultation paper:

Raising the age [of sexual consent] would provide children and young people with an additional measure of protection until they reach a higher level of maturity and understanding about the issues involved in engaging in sexual activity. It would be more consistent with the treatment of children in other activities, such as leaving school, driving and even getting married (Department of Justice 1999:8).

¹ Department of Justice, Canada “Child Victims and the Criminal Justice System: A Consultation Paper” (November 1999).
The view that fourteen and fifteen year olds need protection due to their lack of maturity is not necessarily shared in other sectors. In the youth justice system for instance, rather than protection, the focus is on the responsibility of youth for criminal activity. As a result, young people over the age of fourteen are seen as mature enough to be held accountable in the adult justice system under the transfer provisions of the *Young Offenders Act*. Treating fourteen year olds as adults in this context means facing adult penalties (e.g. a sentence of life in prison), and in certain cases, serving the sentence in an adult facility.

Ontario’s policies concerning welfare benefits provide another interesting example of the contradictions inherent in governing young people in a manner where the importance of age depends on the circumstances at hand. The *General Welfare Act* also takes a protectionist stance in dealing with youthful applicants, but in this case, the focus is on limiting their autonomy outside of the family unit. First, benefits cannot be given to anyone who is under the age of sixteen except in exceptional circumstances.\(^2\) If an individual under the age of sixteen applies for assistance, the Children’s Aid Society or a foster parent or guardian must be called upon to respond to the financial needs of the young applicant.

In the case of sixteen and seventeen year old applicants, the legislation prohibits the provision of welfare assistance unless it can be shown that there are special circumstances for receiving benefits (such as in cases of abuse, parental withdrawal or a parental inability to support the youth). In such cases, the welfare administrator may require a family assessment and verification from a third party in order to confirm that

---

\(^2\) Youths under the age of sixteen are not eligible for benefits in their own right, unless they are sole support parents.
special circumstances do in fact exist. Sixteen and seventeen year old applicants must also have appropriate living arrangements and have regular contact with a ‘responsible adult’ or agency, and they must be registered and attending school or an appropriate training program full time. However, youths who are actively seeking full time work are not considered eligible; they must be in school or a training program. These limitations on assistance apply even for youth who are themselves parents and for youths who apply as a couple (where both are aged sixteen or seventeen and where they can prove to the welfare caseworker that they are genuinely co-habitating as spouses).3

If a young person is successful in receiving welfare assistance, they cannot collect their benefits personally. Rather, “the delivery agent shall only pay assistance to a trustee, guardian or similar party on behalf of a person aged 16 or 17”. Thus, these provisions are premised upon protecting sixteen and seventeen year olds by compelling them to remain within the family unit. And even when successful, they are dependent on a responsible adult to give them their cheque.

The social assistance system should strive to protect those youth who are in need while maintaining the integrity of the family unit. It should not contribute to the breakup of the family unit. Neither should the system be perceived as providing an economic incentive to the youth to leave home or to his or her family to support the youth doing so. This means that actions directed towards reconciliation should be considered throughout the assessment process (Ministry of Community and Social Services 1998: 6).

In contrast, the transfer provisions within the Young Offenders Act see sixteen and seventeen year olds who commit serious offences as individuals who need to be held fully accountable for their criminal offences. The YOA makes specific reference to the fact that sixteen and seventeen year old offenders, who commit a range of serious violent

---

offences, will presumptively be transferred to adult court and face adult penalties
(including life imprisonment), unless they can provide a case for remaining in youth court
and benefitting from their 'youthful status'. Moreover, the Ontario conservative
government which was responsible for the above noted welfare provisions, specifically
prohibiting sixteen and seventeen years old recipients from collecting their welfare
cheques personally\(^4\) has been lobbying the Federal government to lower the maximum
age of criminal responsibility so that sixteen and seventeen year olds are dealt with as
adult offenders in all cases of law-breaking\(^5\).

With these contradictions in mind regarding the regulation of young people based
upon age, the following study examines decision-making in the context of youth justice.
Fundamentally, this study reveals that age and apparent maturity appear not to be related
to decisions made about youthful offenders. However, as will become apparent in the
following chapters, 'youthful' status was recognized in the courts when it came to the
level of supervision that was necessary to manage youths. In cases where the youth was
released into the community, there were significant differences in handing down
'curfews' and 'boundary conditions' for younger youths versus their older counterparts.
Thus, depending upon the situation, the courts do seem to concern themselves with age,
but only in relation to the level of supervision that 'younger' youths need to have
available to them.

Therefore, within the youth justice system there are different logics of regulation
when dealing with adolescents. While legal controls or prohibitions are placed on orders
of release for some young people for their own protection, others in the same age group

\(^4\) Bill C-142 Social Assistance Reform Act (1997).
are constructed as responsible actors, where the emphasis is on their offending behaviour rather than their lack of maturity or state of dependency. It appears that when it comes to young people who, among other situations in their lives have also committed an offence, the message is clear -- childhood ends when you violate the law.⁶

Thus, in seeking to understand the social construction of adolescence in the criminal justice system, my analysis takes the position that there are two irreconcilable principles underlying the way in which young people in conflict with law are regulated. There is a concern for protecting and guiding young offenders due to their age and state of dependency. At the same time, the predominantly adversarial system in which the youth justice system operates is based upon responsibility and accountability for one’s actions with a recognition of youths’ state of dependency. This underlying tension may account for the fact, as noted by others, that the youth justice system is moving more and more in the direction of the adult criminal court model which cannot accommodate both protection and responsibility (Trépanier 1999, Feld 1997). Particularly serious cases, at times, serve as an impetus for legislative change in the direction of an adult, more punitive, accountability-driven model for young offenders (Trépanier 1999).

The following account of decision-making about young offenders by court practitioners and members of the public reveals the arbitrary nature and fluidity of categories of age, but it also allows us to explore how ‘youth’ plays out in the youth justice system.

---

⁶ Hunter Hurst, Director, National Centre for Juvenile Justice (Annual Report 1995 p.3).
Present legal context:

As with most western industrialized countries, the Canadian legal system maintains three discrete age groupings with differing levels of criminal responsibility. Generally, every person who commits a crime in this country falls into one of three groups. There are those offenders whom society sees as having no criminal responsibility (in Canada, children under the age of twelve). Then there are offenders who have mitigated criminal responsibility because of their special status as 'youth' (those between the ages of twelve and eighteen in Canada), and finally there are offenders who are fully responsible for their crimes (adults).

The middle group, represented by the youth justice system, has been characterized as a ‘modified justice model' (Corrado 1992) where youth are responsible for their actions, but are not as fully responsible as adults due to their lack of maturity and development and the understanding that adolescence is a period of transition from childhood to adulthood. The governing body of legislation for young offenders, the *Young Offenders Act*, makes explicit reference to both 'chronological age' and 'maturity and dependency' as indicators of unique treatment apart from adults under the law. First, chronological age is referred to within the *YOA* to demarcate differing levels of criminal responsibility at the ages of twelve, fourteen, sixteen and eighteen. For example, the age of twelve is the minimum age in which young people can be held legally responsible under the *YOA*. Only youth who are fourteen and older can be legally transferred to adult court. A youth over sixteen may be presumptively transferred to adult court in serious cases, and at age eighteen, offenders are dealt with as adults in ordinary court. Thus, this series of legal transitions based upon chronological age would appear to see adolescence
as a slow transition or continuum from childhood to adulthood in terms of criminal responsibility.

At the same time, Sections 3(a.1) and 3(c) of the YOA, which are contained within the “Declaration of Principle”, provide guidance to decision-makers in applying the law to all young people between the ages of twelve and eighteen who commit offences. As these two principles put forward, young people who commit offences are accountable, but not to the same degree as adults. Furthermore, young people who break the law require supervision, discipline and control but also, as a result of their state of dependency and maturity, they have special needs and require guidance and assistance. This approach would appear to see adolescence as a homogenous group where mitigated responsibility applies unitarily to all youths under the law.

The question that arises from these differing conceptions of youthful responsibility under the YOA relates to the way in which the concept of ‘youthfulness’ practically affects decision-making. Is the period of life between the ages of twelve and eighteen conceived of as a slow developmental transition to adulthood, or is it a unitary transition from adolescence to adulthood where developmental differences are taken into account? The language of the YOA does not appear to answer how decision-makers should interpret these questions. As stated above, chronological age is only explicitly referred to in the case of transfer to ordinary court in the present legislation, while s.3 the Declaration of Principle is meant to apply to all youth. The language in Section 3 changes from a concern with welfare and protection to a discussion on responsibility and accountability throughout the statement. Thus, even from the Declaration of Principle

---

7 Although in the original YOA, there was a provision relating to the use of secure custody for those fourteen and under s.24 (3) YOA.
which is meant to guide practitioners in applying the Act, we can see that ‘youthfulness’ is not always the critical factor in dealing with young offenders. The language of the Young Offenders Act embraces the co-existence of incongruent discourses of protection and responsibilization in order to construct youths as either fully responsible or as misguided children depending upon the circumstance. Part of this may be attributed to the wider social climate of youth justice. Public opinion and policy changes in the youth justice system over the past ten years have focused on holding young offenders more accountable for their actions and moving the system closer and closer in the direction of the adult criminal justice system. At the same time, although to a more limited extent, we find the more welfare-based discourse of ‘misguided youth’ with special needs continues to permeate discussions about the treatment of young offenders.

The purpose of this thesis is to explore this matter by addressing two main questions through empirical analysis. First, what role does chronological age or apparent maturity play in court proceedings for youth and public responses to youth within the ‘transition’ period from childhood to adulthood? And second, how does decision making about youth relate to the principles and the legislative provisions of the Young Offenders Act? The implications of the findings will be discussed in the concluding chapter.

Historical context and the process of reform

The idea of differential treatment in law for young offenders is a relatively recent phenomenon. The first criminal law dealing with youth as a separate legal category was the Juvenile Delinquents Act in 1908. Prior to this, children of any age were subject to the criminal law (Griffiths and Verdun-Jones 1994) although separate custodial facilities
for neglected and delinquent youth did exist prior to the passing of formal legislation (Trépanier 1999, Splane 1965). There is also evidence that even at a time when youth were not seen as a separate legal category, ‘youthfulness’ was considered to be a mitigating factor in court decision-making in other jurisdictions (Smundych 1995, King 1984).

Despite the use of informal discretion for younger criminals, English common law’s doli incapax rule brought forth a more formal recognition of youthful status and was applied to the first Criminal Code in Canada in 1892 (Reid 1986, Dalby 1986). The doli incapax rule maintained that persons under seven years old were deemed to be legally incapable of committing an offence. There was a presumption of doli incapax for those aged seven to fourteen years old, but the prosecutor could contest this. Anyone fourteen years and older was fully responsible for their actions.

Changes in the nineteenth century brought on by the industrial revolution resulted in a re-examination of the place that young people held in society (Tanner 1996). Part of this was as a result of the large numbers of youth in the general population. A census taken of Upper Canada in the mid-nineteenth century revealed that over one-half of the population of the province was under sixteen years of age (Splane 1965: 220). In addition, progressive reformers were influential in providing the impetus for welfare institutions, compulsory education, and a multitude of ‘experts’ who defined the category of adolescence as an objective arena for social study (Iacovetta 1996). These changes not only affected the place of the child in society but also those who would care for children.

As Rothman (1980) points out, an American psychologist by the name of G. Stanley

---

8 The age of seven was thought to coincide with Roman Law and the presumed onset of puberty. The actual concept of “Infantia” meant “incapacity to speak out”. (Fox 1984, Reid 1986).
Hall, whose interests centred around child development and evolutionary theory, had a large influence on how ‘motherhood’ was to be deployed. Hall put forward the thesis that childhood was composed of very distinct stages that required particular responses by caregivers and, as a result, mothers needed to be trained to be appropriately responsive to the complexity and stages of child development. In addition, a changing perception regarding the causes of delinquency in the late nineteenth century brought forward ideas about protecting society through protecting children. These ideas became central to the debates preceding the passing of the Juvenile Delinquents Act (Trépanier 1991). Bad homes, unhealthy child rearing and hereditary influences were all thought to contribute to the general problems facing youth (Iacovetta 1996).

The Juvenile Delinquents Act was passed in 1908 with a debate that lasted for about one hour in Parliament (Corrado and Markwart 1992). The new legislation was premised on the idea of treating and protecting children and, thereby, preventing criminality. There was no distinction made between neglected and delinquent children “since the former were but incipient versions of the latter” (Fox 1984: 152). The welfare approach of the JDA was premised on the ‘best interests’ of the youth suggesting that the state protected youths in every respect making the need for individual due process rights incidental. State authority represented, as closely as possible, the relationship between parent and child (Rothman 1980, Simon 1995, Chunn 1992) and workers of the court were able to reproduce as closely as possible what they considered to be ‘good parenting’ and a ‘sound’ upbringing.

Despite the benign intentions of the Act, state intervention without due consideration for the rights of young people resulted in challenges to the constitutionality
of this approach in the United States. However, when early attacks were made in this regard against juvenile delinquency statutes in the U.S., the *parens patriae* doctrine was sufficient as a justification for intervention on wider welfare grounds (Fox 1984). As a result, the juvenile court operated for the greater part of the century in Canada and the United States without procedural protections for youth. There were no legislative guidelines governing judicial sentencing, there were few lawyers representing young people at court, the rules and proceedings were relaxed and informal, and judgements were shaped by individual diagnoses often through the taking of statements or ‘confessions’ as integral to the rehabilitative process (Simon 1995, Corrado and Markwart 1992). Moreover, juvenile courts defined and designated other social agencies that previously did not have the power to intervene in young people’s lives to have jurisdiction over a wider range of young people (Simon 1995). Thus, “doing good” took precedence over any interest in legal rights (Chinn 1992), and the ideology of treatment was so powerful that “. . . there were occasions when guilt seemed to be presumed so that “treatment” would not be delayed by “unnecessary formalities” (Bala 1992: 22).

There was, however, skepticism even within the first years of passing the juvenile delinquency legislation concerning the view that legal protections did not need to be extended to youths (Fox 1984). These concerns became more emphatic in the early part of the 1960’s when academics and political commentators voiced concerns that the juvenile court was operating as an ordinary criminal court without the safety of due process protections, nor the security of ‘best interests’ (Bala 1992).

The philosophical tension between state intervention justified on the best interests of youth and the potential extension of political rights to young people began to present a
problem for youth justice legislators and reformers of the system during the 1960’s. The political climate at the time was one in which civil rights concerns were at the forefront in both the United States in Canada, and a number of influential U.S court cases\(^9\) provided fuel for the changing sentiment about the place of legal rights in youth justice. In Canada, juvenile justice reform began through a predominantly bureaucratic initiative.\(^10\) The appointment of an advisory committee by the Department of Justice in 1961 provided the first step in formally rethinking juvenile justice. Recommendations, which came out in 1965, included provisions for respecting the legal rights of young people before the courts as well as providing resources to adequately facilitate rehabilitation. But in order to create a proper ‘space’ for the extension of legal rights for youth there would have to be a fundamental shift in philosophy away from a *parens patriae* approach. Despite this, while the Department of Justice Committee endorsed respect for the legal rights of youth, they still wanted to maintain the recognition of rights within the overall welfare/treatment model of the *JDA*.

...the Committee seemed to be saying that a Welfare model, circumscribed by a due regard for legal rights, was still the proper course if only the juvenile justice system could be given the resources to implement some of the idealistic goals of rehabilitation. In effect, rehabilitation had not been given a chance (Corrado and Markwart 1992: 148).

Consequently, the ideal of the welfare based juvenile court maintained a strong position even within a changing paradigm that began to recognize the need for extending legal rights to youth.

After a number of draft bills, in 1970, Bill C-192, the *Young Offenders Act* was introduced by the Liberal government. However the position taken in Bill C-192, which


\(^{10}\) See Corrado and Markwart 1992 for a detailed explanation of the process of juvenile justice reform in Canada.
tried to mix a welfare approach with a legalistic approach, failed due to a strong opposition by welfare/treatment interest groups who considered the bill to be too legalistic and punitive. But as Corrado and Markwart (1992) point out, only a decade later, in 1982, the philosophy of the welfare approach was almost completely abandoned in favour of the crime control and justice model principles of the Young Offenders Act, which was implemented in 1984. The question raised is why welfare interest groups were rendered silent in the following years of debate.

Part of the reason for the abandonment of a welfare based approach for dealing with young offenders was the influence of the Charter of Rights and Freedoms enacted in 1982. An extension of the principles of due process, legal rights, and procedural fairness had to be incorporated into new legislation since differential treatment based upon age violated young people's newly defined constitutional rights. As well, skepticism about the true principles of the juvenile court being implemented as promised became a concern in academic and political circles. Social science research had lasting effects on policy. For instance the ideas of labelling theory and 'radical non-intervention' had far-reaching consequences on juvenile justice policy resulting in the endorsement of programs to divert young offenders away from formal court processing. At the same time the highly quoted line that 'nothing works' provided the fuel for a scathing critique of the principles of rehabilitation in the justice system and provided a reason to re-think the fundamental objectives of penal policy.

In 1973 another committee was established by the Solicitor General to further investigate how to approach juvenile justice policy, and in 1975, a report called Young

---

11 E. Schur 1973, E. LeMert 1972
*Persons in Conflict with the Law* (YPICL) which contained new draft legislation was the result. This report put forth a legal rights orientation while retaining some aspects of the welfare model in an attempt to arrive at a politically acceptable compromise to accommodate a wide range of interest groups. In addition, the report recommended the raising of the criminal age of responsibility from seven to fourteen years and setting a uniform maximum age at eighteen years across the country, which would have excluded a large number of children from adult criminal control (Fox 1984). When this report was discussed, there were few objections to the philosophical orientation of the document, but rather matters relating to the division of powers between federal and provincial governments and issues relating to the costs of implementation were of concern. Specifically, provinces were opposed to the costs associated with raising the maximum age of criminal responsibility to eighteen years.\(^{13}\)

A few years after the ‘YPICL’ report, the Liberals released a document outlining the government’s position on young offenders in conflict with the law. With respect to the controversial age boundaries, the Liberals proposed a lower minimum age of criminal responsibility than was proposed in YPICL.

Under the proposed Young Offenders Act the minimum age of criminal responsibility would be twelve years. In setting this age, consideration has been given to the state of development of the child in physiological, mental and emotional terms, particularly as these factors apply to the formulation of a criminal intent. The setting of a precise age is necessarily arbitrary as children vary greatly in their rate of development but it is assumed that deviant behaviour by children under the age of twelve is better and more effectively dealt with under provincial legislation pertaining to child welfare or youth protection (Solicitor General Canada 1977: 5).

\(^{13}\) Under the *Juvenile Delinquents Act* the minimum age was 7 years and each province had the discretion over the maximum age of criminal responsibility. Some provinces had a maximum age of 18 years (e.g., Quebec) while others retained a maximum age of 16 years (e.g., Ontario).
A uniform maximum age across the country was also put forward. While the 1977 report recommended that this be eighteen years, the report maintained that the provinces would still have the ability to set the maximum age lower (at age sixteen or seventeen) if they could not all agree on eighteen.

After the Liberal government was defeated in 1979, a newly elected Conservative government put forward their own proposals for new youth justice legislation in response to provincial concerns but kept the general philosophical framework of the Liberal proposals intact. However, it also incorporated the notion of the 'protection of society' as a key consideration. The minimum age jurisdiction in this report was kept at twelve, while the maximum age was lowered from the earlier Liberal report to sixteen. But once again, the Conservative proposals were prepared to consider a uniform maximum age at seventeen or eighteen if there was an agreement among provinces to this effect (Solicitor General Canada 1979:5). When the Liberal government came back into power in 1981 they introduced legislation which was nearly identical to the 1977 proposals. Bill C-61, *The Young Offenders Act* continued to be criticized for, among other things, the mandatory maximum age provisions (Corrado 1992).

The proposed *YOA* was the subject of intensive study and debate in Parliament (Corrado and Markwart 1992). More than 40 interest groups from diverse perspectives made presentations to the parliamentary subcommittee, and after a lengthy process of consultation and debate, Bill C-61 the *Young Offenders Act*, received all party support in 1982 and came into force on April 2, 1984 (Bala 1992). The implementation of the new uniform maximum age came in 1985 in order to accommodate those provinces with a former maximum age of sixteen (Corrado and Markwart 1992). The form that the *YOA*
took was influenced by a climate of general disillusionment with the welfare model that moved the focus towards what the system could actually accomplish. "If the juvenile justice system could not "do good", it could at least "do justice" (Corrado and Markwart 1992: 155).

The current Young Offenders Act still retains the mix of welfare and justice models in attempting to deal with all potential contingencies in dealing with youth criminal behaviour. However, since its inception, three separate sets of amendments to the YOA\(^{14}\) have moved the Act towards a greater crime control orientation. Bill C-106 in 1986 retained the applicable age of transfer to be fourteen years old, which was similar to the original Young Offenders Act and the former Juvenile Delinquents Act. Thus, even for serious offences, there was a preservation of a homogenized view of the category of 'youth'. However, only 9 years later, Bill C-37's presumptive transfer changes specified differences among youth in terms of transfer. Most notably, sixteen and seventeen year old youths who committed specific serious violent offences would be presumptively transferred to adult court. Currently, Bill C-3, the proposed Youth Criminal Justice Act which is being put forth to replace the YOA, will among other things, lower the age of presumptive adult sentences to fourteen years old for certain serious violent offences similar to the Bill C-37 offences (such as murder, manslaughter, aggravated sexual assault). At the same time however, Bill C-3 encourages the use of a full range of community-based sentences and effective alternatives to the justice system for youth who commit non-violent offences.

As stated in the Department of Justice news release:

The Youth Criminal Justice Act is based on an accountability framework that promotes consequences for crime that are proportionate to the seriousness of the offence. More serious offenders could receive adult sentences or sentences of custody. Less serious offenders will be dealt with through measures outside the court process or be subject to constructive community-based sentences or alternatives. The Act emphasizes that, in all cases, youth should face consequences that promote responsibility and accountability to the victim and the community and teach good values by helping the young person understand the effect of his or her actions.

Thus, it appears that the new legislation is moving in a direction which more explicitly promotes proportionality in sentencing based upon the offence. Very serious offenders may receive more adult-like treatment, but measures outside of the formal court process are reserved for less serious young offenders.

Developmental Research

While changes have occurred and are still occurring in youth justice legislation towards a crime control orientation, the view of adolescents as having mitigated responsibility for their criminal activities has not been set aside. As stated earlier, the current legislation is explicit in referring to both ‘chronological age’ and ‘maturity and dependency’ as indicators of unique treatment apart from adults under the law. Chronological age is referred to within the YOA to demarcate differing levels of criminal responsibility at different ages. The age of twelve, fourteen, sixteen, and eighteen are all points of transition in terms of criminal responsibility, and these gradations in chronological age appear to reflect a view of adolescence as a slow transition from childhood to adulthood. There is a fair amount of research to draw upon which supports this conception of adolescence as a slow transitional period.
For example, recent research in developmental psychology has pointed out that there are significant differences between younger and older adolescents in their understanding of the legal process. A review of this research by Scott and Grisso (1997) has revealed that youths under the age of fourteen differ (and are at a disadvantage) compared to their older counterparts in terms of their ability to competently participate in their own legal defense, and in their understanding of the legal process and its basic purposes. The authors also point out that youths between ten and thirteen years of age were significantly less likely to think “strategically” about pleading decisions and appeared to be less capable of imagining risky consequences during hypothetical problem solving situations. Younger children were also likely to have considered a more constricted number and range of consequences (Scott and Grisso 1997: 170-171).

Canadian research on youths’ understanding of the legal process has found similar differences across age groups. For instance a study by Abramovitch, Higgins-Biss and Biss (1993) found significant differences among youth in different age groups in terms of their ability to understand basic legal rights on questioning, the right to counsel and the right to be provided a lawyer free of charge. Fewer of the ‘younger youth’ (ages 10.50 years to 12.92 years) comprehended these rights as compared to those in the older range of (16.58 to 19.92 years). Peterson-Badali and Koegl (1999) also found significant differences among age groups in terms of general knowledge of the YOA. Younger youths were different from older youths in that they were less accurate in their knowledge of various aspects of the youth justice system such as minimum and maximum ages, minimum ages for transfer to adult court and the use of pre-trial detention.
Bartusch, Lynam, Moffitt and Silva’s (1997) examination of developmental
theory in explaining different kinds of antisocial behaviour, found that the causes for
identical types of antisocial behaviour differed at childhood (ages 5, 7, 9, 11)\(^{15}\) as
compared to adolescence (ages 13, 15). Their analysis showed that childhood antisocial
behaviour was related more strongly than adolescent antisocial behaviour to low verbal
activity, hyperactivity, and a negative/impulsive personality whereas; adolescent
antisocial behaviour was associated with peer delinquency. As well, childhood antisocial
behaviour was more strongly associated with convictions for violence while adolescent
antisocial behaviour was more strongly related with convictions for non-violent offences
(Bartusch, Lynam, Moffitt and Silva 1997).

Tonry (1999) points out that there are also differences in the nature of offending
depending on the age of the youth. The prevalence of offending peaks in the mid-teens
for property offending, and in the late teens for violent offences. Moyer (1996) also
found evidence to support differences among youth in the nature of offending. Using the
revised uniform crime report data for selected jurisdictions in Canada in 1992-93, Moyer
found that police suspects who were twelve and thirteen years old were more likely to be
apprehended for minor property offences, such as theft under and mischief or vandalism,
than were any other age group (Moyer 1996: 36). In relation to the victim-suspect
relationship, the older the suspect, the greater was the likelihood that the victim of an
offence was a stranger. Of those suspects who were twelve and thirteen years old, 80% of
victims were either friends or family. For sixteen and seventeen year olds 65% of
victims were friends or family of the suspect (Moyer 1996: 44). There were also
differences in the use of ‘weapons’ based upon age. Moyer found that the use of physical

\(^{15}\) Although youth in this age group do not fall within the jurisdiction of the YOA
force decreases with age and the use of 'other' weapons increases with age. As well, major injuries increase by age of the suspect although these percentages were quite small in every age category (Moyer 1996: 45-48).

It is also important to point out that the intersection of age and gender are significant factors in understanding differences in youthful offending. For example, the participation of male youths in criminal activity appears to increase with age, whereas for females there is a decrease in involvement in criminal activity at about the age of sixteen (The Daily, Canadian Centre for Justice Statistics 1998). Furthermore, in an examination of violent youth crime statistics across Canada in 1998, it was found that female youths charged with violent crime tended to be younger than their male counterparts. Among male youths, the violent crime rate increased gradually with age, with the highest rate being among 17-year-olds. Among females, the rate peaked at the ages of 14 and 15 (Juristat 1999).

Thus, this body of research suggests that there are important differences among youthful offenders in terms of age and criminal activity. These differences do not support a view of youth as a homogenous group, rather, they provide support for the contention that youth is a slow transition from childhood to adulthood, and consequently, one would expect to see differences in the processing of 'younger' youths (closer to the minimum age of responsibility) and 'older' youths who are closer to adulthood.

The following chapters will attempt to investigate how chronological age and apparent maturity relate to decisions made about young people in conflict with the law. Chapters 2 and 3 will provide an empirical examination of the ways in which the principles of the YOA and the apparent maturity of a youth affect decisions made in bail
and sentencing cases. Chapter 4 explores public responses to a young offender constructed on varying dimensions of 'maturity' and how this affects sentencing preferences and offender ratings. Chapter 5 presents a summary of key findings and will explore in more detail the implications of these findings for the treatment of youths in the youth justice system as well as how these findings may provide us more insight in understanding the contradictions in how adolescence is constructed depending upon the context.
Chapter Two – Bail hearings

Introduction

The purpose of this chapter is to present findings on the way in which youths charged with offences under the Young Offenders Act are constructed at their bail hearings. Bail (or pre-trial detention) hearings are one of the ‘sites’ in which there is a process of describing young people accused of offences in a particular way by Crown and defense attorneys, in order to make a case for further detention or for release. As such, examining the information presented in bail hearings can provide us with some insight into this process of describing young persons accused of offences, and what this means for how a youth comes to be understood in court. Integral to this analysis is the construct of age as a factor in decision-making in youth court bail hearings. Thus, both contextual information relating to age, and the actual chronological age of youths in this sample of cases, will be examined in order to determine the effect of age on decision-making.

There is a paucity of research on pre-trial detention hearings for youth and even less information on how youth are described during their bail hearings. We know very little about the kinds of information about youth that are seen as relevant to the hearing, nor the effect that these kinds of information have on the outcome of the hearing. Some of what we know comes from media sources, but these descriptions often describe youth in detention as taking pride in their delinquent image. As one article put it “...[youth in court] have developed a tough veneer. For them, being detained is a kind of warped rite of passage... (h)andcuffs are a badge of honour.”

---

1 Black, Debra “A Court Just for Youths” The Toronto Star April 5th, 1998.
This chapter will examine how youth are constructed in their bail hearings by exploring the effect of legal variables, factors related to the support available to a youth, and variables relating to how the youth ‘appeared’ in court. More specifically this chapter will present findings on how the ‘youthfulness’ of an accused, captured by information observed in court, and the youth’s chronological age, related to the decision-making process in bail hearings.

Legal Context

The legal framework for youth bail hearings consists of four principal bodies of legislation. The main law, which governs bail hearings, is the Canadian Criminal Code section 515, which provides the rules for judicial interim release for all offenders – adult and youth. Thus, section 515 is written without specific reference or rules for young people who commit offences.

Generally after a decision is made by a police officer to detain a youth after an arrest (governed by section 495 of the CCC), the law holds that criminal justice agents have 24 hours, or as soon as is practicable, to place the accused in front of a justice of the peace or judge for a bail hearing. At the bail hearing, there are three grounds for which bail can be denied. The first is to ensure the attendance of the accused person for trial, and the second is that detention is necessary for the protection of the public. The third, recently added tertiary ground, justifies detention if it can be demonstrated that granting release would erode public confidence and bring the administration of justice into disrepute. Before the tertiary ground was added to the Criminal Code, the grounds for detention were to be considered in order. However, research suggests that even when
judges or justices were to consider these grounds in order, they did not always do so, but rather sometimes went directly to the condition that justified detention (Gandy 1992).

There are a number of different ways that an offender can be released on bail. The accused can give an undertaking to the justice that he/she will show up for trial and this undertaking can be placed on the offender with or without conditions. The offender can also be released on his/her own recognizance with or without sureties or monetary deposit and with the possibility of conditions being placed on release. If released the kinds of conditions that can be placed on an accused are set out in section 515(4) of the *Criminal Code*:

- to report to a peace officer or designated person at certain times.  
- to remain within a certain jurisdiction.  
- to notify a peace officer of a change in address or employment.  
- to abstain from communicating with any witnesses and refrain from going to any place expressly named in the order.  
- to deposit one's passport.  
- and to comply with any other reasonable conditions that the justice considers desirable.

For youths, some of the conditions placed on them fall into the last category – 'other reasonable conditions that the justice considers desirable.' While not made explicit in the legislation, these conditions are sometimes specific to age, and appear to relate to the availability of informal structure and supervision. For example, having a curfew, obeying house rules, and attending school are all conditions that are routinely placed upon accused youth as part of their orders for release. (A further examination of the implications of this will be discussed later in the chapter).

The second main body of legislation that influences youth bail hearings is the *Young Offenders Act*. While the *YOA* does not have it's own set of rules governing

---

2 A tertiary ground for detention was legislated part way through the collection of these data.
judicial interim release, there are a few sections which are to be considered only in youth pre-trial detention hearings. First is section 7.1 of the YOA - the placement of the youth into the care of a responsible person. This section states that if, at the outcome of a pre-trial detention hearing it is deemed that the young person warrants an order of detention and there is a ‘responsible person’ willing and able to take care of and exercise control over the young person, and the youth is willing, the justice can order the youth to be placed in the care of this responsible person who takes a written undertaking to comply with this arrangement. In practice, the ‘responsible person’ provision is almost never used (Federal-Provincial-Territorial task force report 1996; Gandy 1992). This finding is consistent with my observation – this provision was never raised in court in any of the 118 cases observed as part of this study.

The other aspect of the YOA that should be considered in youth bail hearings is Section 3 of the YOA - the Declaration of Principle. The Declaration sets out the framework for understanding and dealing with young people who come in contact with the law, and attempts to provide guidance to criminal justice agents in dealing with youth who commit offences. There are four main themes found in the Declaration of Principle: prevention, legal rights of youth, mitigated responsibility & special needs of youth, and the protection of the public. The results from this study showed that reference to the YOA principles in bail court was a rarity. In only 3 cases out of the 118 observed, was any

---

3 If a responsible person does come forward to supervise the youth and willfully does not fulfill this responsibility, the ‘responsible person’ can be found guilty of an offence punishable on summary conviction (S 7.2 Young Offenders Act).
4 the newly proposed Bill C-3 the Youth Criminal Justice Act has amended this section so that the responsible person provision specific to youths might be used more often. Section 31(2) states that “If a young person would, in the absence of a responsible person, be detained in custody, the youth justice court or the justice shall inquire as to the availability of a responsible person and whether or not the young person is willing to be placed in that person’s care.”
reference made by anyone (accused’s counsel, Crown or justice) to the principles of the YOA.

Third, the *Charter of Rights and Freedoms* is also relevant to youth bail hearings. Among other rights, accused persons in bail hearings have:

- the right not to be arbitrarily detained or imprisoned,
- the right to be presumed innocent until proven guilty according to law in a fair and public hearing and by an independent and impartial tribunal,
- the right not to be denied reasonable bail without just cause,
- and the right to equal protection and equal benefit of the law without discrimination and in particular without discrimination based upon...age.

Finally, the fourth major piece of legislation pertaining to youth court bail hearings in Ontario is the *Child and Family Services Act*. Each province has its own child welfare legislation which governs the administration of justice for young people. Section 93 of the Ontario *Child and Family Services Act* sets out that factors to be considered in determining the level of temporary detention upon being arrested by a police officer.

The common presumption behind all of these pieces of legislation is that in all possible cases, the *least intrusive measure* should be taken. For example, at the first stage of proceedings, there is a duty on the police officer not to arrest, if a release by way of summons or appearance notice is sufficient. Similarly, at the bail hearing, a justice or judge has a duty to release without conditions unless a case can be made for doing otherwise. The Ontario *Child and Family Services Act* maintains that open custody should be presumptively used for temporary detention, unless it is necessary to ensure the youth attends court or there is a threat to public safety. Finally, the Declaration of Principle in the YOA states that young offenders have the right to the least possible interference with their freedom that is consistent with the protection and safety of the public.
The Decision-Makers:

The cases that were observed in this study were the product of a ‘filtering process’ of decisions made by various criminal justice personnel. Of primary importance is the fact that these are cases where an offender was apprehended by police. But one must keep in mind that a large amount of criminal activity goes unreported (Doob et al 1995; Griffiths & Verdun-Jones 1994). Second, these are cases where the decision to arrest was made and a decision to detain was seen as being in the public interest by both the arresting officer and the officer in charge (Sec.498 CCC). Thus, what comes into court is highly dependent upon a series of discretionary decisions made by the police. And as other research suggests, police decision-making at arrest is significantly related to both legal factors (number of previous contacts with police, seriousness of the offence, prior record of the accused) and non-legal factors (race, gender, socio-economic status, situational factors) (Wortley & Kellough 1998, Sealock & Simpson 1998, Wordes, Bynum & Corley 1994, Doob & Chan 1982, Carrington, Moyer & Kopelman 1988).

The next phase of the ‘filtering’ process is the decision made by the Crown attorney. After reviewing the case, the Crown may seek to contest bail release or may consent to the release of the youth, with or without conditions. This decision, as we shall see, is pivotal to the outcome of the case (Wortley & Kellough 1998; Hucklesby 1998).

Finally, the justice of the peace or judge listens to evidence presented at the bail hearing. Based on the testimony of sureties, parents, and at times the accused youth, a decision is made either to detain the youth, or release the accused with or without conditions. Thus, this last phase of the process — the bail hearing — is highly dependent upon earlier decisions made which brings the case to court.
Research Methods

From June to the end of August 1997, a research assistant and I observed bail hearings at 4 different youth courts in the Toronto area. The majority of cases (61.9%) were observed in Canada’s largest youth court, which deals only with young offender cases, and also hears all of the city’s drug cases involving youth.

The data collection form used to collect information in court included those factors thought to be associated with youth detention in previous research (see Carrington et al 1988 for a review). As well, other more subtle pieces of information about the way the youth may have been perceived in court were also recorded. For each case, therefore, we recorded demographic information (age, gender); legal factors (current charge, prior record); social factors (living arrangements, school involvement) and personal characteristics (how the young person was dressed, how the youth acted in court) in the bail hearings. I was interested in examining the impact of the information which was available in the courtroom to various criminal justice personnel, in particular the judge or justice, who would be making the final decision. In total, 118 cases were observed which resulted in bail being either granted or denied.

The theory behind collecting data by courtroom observation to examine bail decision-making is that the information presented in the courtroom is almost as complete as the information that the presiding justice of the peace or judge has in making the bail decision. Generally, the information that justices or judges have before them is the youth’s prior record information, outstanding charges, and the current charges before the court. In most cases, we had the court dockets which contained the information on current

---

5 See Appendix B for inter-rater reliability
6 See Appendix A
charges. And, in many cases, the Crown read out prior record information in court.
Consequently through court observation, we were able, as nearly as possible, to ‘see what
the decision makers saw’.

Given the limited amount of written information that a judge or justice relies on to
make bail decisions, the information presented and seen in court is quite important in how
the youth is constructed in the bail hearing and is thus likely to affect the decision-making
process. Conversely, what is not mentioned is equally important in how the youth is
understood in court during the bail hearing. I was interested in examining what this means
for the ways in which youth are constructed in court, and more specifically, how the
category of ‘age’ fared in the decision-making equation.

While this method of data collection allowed me to obtain a close proxy of judicial
decision-making, there were some difficulties in interpreting this information. Due to the
consensual nature of the court proceedings, there was not as much information brought
out in court as was expected and this had implications for the analysis. As will become
apparent in the following chapter (as well as Chapter 3 on sentencing hearings), there
were two different ways in which ‘missing data’ were treated. First, there were data that
were actually missing and thus were excluded from statistical analysis since they were not
coded by the research assistant or me. Therefore the total number of cases in the following
tables may vary due to the exclusion of missing data. The other kind of ‘missing data’ was
information that was not mentioned in the court proceeding. I was interested in
exploring how the lack of information in a case might construct it differently. Thus, when
information was ‘not mentioned’ it was also included in the analysis. However, in all
cases, statistics are provided which include cases where information is ‘not mentioned’ as well as statistics excluding the ‘not mentioned’ cases.

This chapter will be divided into 4 sections. The first will present findings on factors related to the Crown’s decision to contest or consent to release. The second section will examine only those cases where the Crown contested release in order to explore decision-making in the bail hearing above and beyond the Crown decision. The third section will look at the overall detention decision by the court which is presumably a combination of the first two decisions. The final section will provide an analysis of bail conditions placed on cases that were released by the court, in order to understand how these conditions may relate to the youth’s age and maturity.

The Sample of Observed Cases

Almost one-third (29.7%) of the accused youths whose hearings were observed in this study were denied bail. Just over half of these young people (50.8%) had as their principal charge® either a violent offence or a charge of break and enter—two offence categories which are likely to be considered to be the most serious by the public and the judiciary. Over half (58.5%) of the sample was 16 years of age or over. The majority of the youths in these bail hearings were male. Females comprised only 17.8% of the cases.

7 The principal charge was computed by coding the first four charges in a case into 6 categories of offences; violence, drugs, break and enter, other property, other Criminal Code and YOA offences (in many observed cases though, there were fewer than four charges). Then, the principal charge was calculated by using the order of categories of offence (consistent with Canadian Centre for Justice Statistics-as above) where the ‘most serious’ category trumped all other charge categories.
Table 1: Comparison of the Sample of Youth Bail Cases observed between June and August 1997 to Cases Heard in Ontario Youth Courts (1997-1998) by Principal Charge, Age & Sex

<table>
<thead>
<tr>
<th>Most Serious Charge</th>
<th>Bail Sample</th>
<th>Ontario</th>
</tr>
</thead>
<tbody>
<tr>
<td>Violence</td>
<td>34.7%</td>
<td>24.5%</td>
</tr>
<tr>
<td>Drug Offences</td>
<td>13.6%</td>
<td>4.3%</td>
</tr>
<tr>
<td>Break &amp; Enter</td>
<td>16.1%</td>
<td>10.6%</td>
</tr>
<tr>
<td>Other Property</td>
<td>20.3%</td>
<td>34.7%</td>
</tr>
<tr>
<td>Other Criminal Code(^8)</td>
<td>11.0%</td>
<td>16.8%</td>
</tr>
<tr>
<td>YOA offences(^9)</td>
<td>4.2%</td>
<td>9%</td>
</tr>
</tbody>
</table>

Age

<table>
<thead>
<tr>
<th>Age</th>
<th>Bail Sample</th>
<th>Ontario</th>
</tr>
</thead>
<tbody>
<tr>
<td>12-15 years old (all offences)</td>
<td>41.5%</td>
<td>51.6%</td>
</tr>
<tr>
<td>16+ years old (all offences)</td>
<td>58.5%</td>
<td>48.3%</td>
</tr>
</tbody>
</table>

Sex

<table>
<thead>
<tr>
<th>Sex</th>
<th>Bail Sample</th>
<th>Ontario</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>82.2%</td>
<td>78.1%</td>
</tr>
<tr>
<td>Female</td>
<td>17.8%</td>
<td>21.9%</td>
</tr>
</tbody>
</table>

Looking at how this sample of cases compares to Ontario youth court cases (Table 1), reveals that the bail sample includes more offences falling into the categories of violence and drug offences than Ontario cases, and a higher proportion of males, and youths 16 years of age or older. This is not surprising since these are youths who were subject to an earlier decision by police to be detained based upon the criteria set out in section 495 of the *Criminal Code*.

Almost all of the youth in the observed sample came into court in handcuffs and in police custody since they had been held in secure detention before their bail hearing (89.8%). Only in 11 cases out of the 118 (9.3%), were the youths placed into an open detention facility before their bail hearing.

\(^8\) such as fail to comply with a disposition, fail to comply with an undertaking (this includes all cases except for YOA offences, drug offence and other federal offences)

\(^9\) such as escape custody, unlawfully at large, fail to comply with a probation order
The factors that are to be considered in determining the level of custody for temporary detention in Ontario are set out in section 93 of the *Child and Family Services Act*.\(^\text{10}\) It is stated that a young person who is detained in a place of temporary detention *shall be detained in a place of open custody* unless a provincial director determines that it is necessary to detain the young person in a place of secure detention to ensure the young person’s attendance in court, or to protect the public interest or safety. Given that the nature of the alleged charges in over 20% of the cases were minor property offences, and that there is a presumption that the least intrusive option -- open custody -- should be used for temporary detention, it is surprising to see how many youth were held in secure detention pending their bail hearing. This however, may be related more to the availability of space rather than criteria set out in the *Act*.

While almost all (89.8%) of the youth in this sample spent some time in secure detention before their bail hearing, most of these youth (70.3%) were subsequently granted bail. The question raised, then, is why are so many accused offenders spending any time locked up before their bail hearing if they are ultimately to be released (also see Gandy 1992)? The negative effects of entering the court in custody or being held in pre-trial detention in relation to the outcome of the case have been well documented. Most studies have concluded that accused persons stand a greater chance of being convicted and receiving a harsher sentence if they enter the court in custody (Friedland 1965, Koza & Doob 1975, Goldkamp 1983, McCarthy 1987, Fagan & Guggenheim 1996).\(^\text{11}\) In fact, as Chapter 3 will show, young persons who were held in pre-trial detention were more likely to receive a custodial disposition than those who were not detained before trial.

\(^\text{10}\) *Child and Family Services Act* chap. C.11, part IV - 1990-91
**Results - The Crown**

In understanding the factors related to bail decision-making in youth courts, the data clearly show that the main decision-maker in youth bail court is the Crown Attorney (Table 2). If the Crown Attorney did not contest release, the youth was certain to be granted bail. However, when the Crown contested the release of the youth, almost 3 out of 4 youths (72.9%) were detained. Although there has been very little information written on the impact of Crown decision-making in youth court bail hearings, these findings are consistent with a study on adult bail hearings which found that the Crown consented to release the accused in 78.4% of cases, and that the Crown’s particular form of release was followed by the judge in 95.4% of the cases where a release order was sought by the Crown (Koza & Doob 1975). Some have argued that this kind of highly autonomous prosecutorial decision-making allows for temporal efficiency and convenience in the courtroom for dealing with overburdened court dockets (Koza & Doob 1975, Klein 1998 in relation to transfer hearings).

**Table 2**

<table>
<thead>
<tr>
<th>was bail granted?</th>
<th>no</th>
<th>yes</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>does Crown contest release?</td>
<td>no</td>
<td>69</td>
<td>69</td>
</tr>
<tr>
<td>yes</td>
<td>35</td>
<td>13</td>
<td>48</td>
</tr>
<tr>
<td>Total</td>
<td>35</td>
<td>82</td>
<td>117</td>
</tr>
</tbody>
</table>

NOTE: Fisher's exact test \( p < .001 \).

---

11 Frazier & Bishop (1985) found that detention status had no effect on the severity of judicial dispositions.
These results reveal the powerful influence of the Crown Attorney and raise questions about the accountability of decision-making in the bail hearing, since a substantial amount of the decision-making control appears to be carried out independently of the information provided at the bail hearing (Hucklesby 1997). In this sense, attempting to look for ‘youth’ in youth court bail hearings may be futile, since contextual information about the youth only occurs within the hearing itself. And as these data show, the Crown is making decisions about release independent of the actual bail hearing.

Section 1 - Legal Factors relating to the Crown Decision

Given the high degree of influence the Crown has in the decision-making process of youth court bail hearings, the first section of this chapter will explore the factors associated with the Crown’s decision to contest or consent to release on bail.\textsuperscript{12}

In order to see if various factors grouped together made a difference in the detention decision, indices were created based upon 3 categories, ‘legal factors’, ‘child vs. adult’, and ‘support factors’ (see Appendix C for a breakdown of the variables included). Each of the variables, if mentioned or observed in the bail hearing, either counted as a ‘point’ toward the overall index category, or if it was a factor which detracted from the overall category - it counted as a negative. Thus, to create the index, the variables were added together and the negative points were subtracted.

Legal Index:

Having a case with a large number of factors of ‘legal seriousness’ was related to the Crown decision to contest bail. As Table 3 shows, as factors of legal serious in a

\textsuperscript{12}Research suggests that the Crown’s decision is highly influenced by the police ‘show cause’ report and recommendations to the Crown (see Wortley & Kellough 1998). Given that this study focused in on court as a site of constructing youth, this information was not accessible.
youth's case increased, so too did the likelihood that the Crown would contest bail release.

Table 3

<table>
<thead>
<tr>
<th>Factors of legal seriousness</th>
<th>no</th>
<th>yes</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 through 2 serious</td>
<td>25</td>
<td>3</td>
<td>28</td>
</tr>
<tr>
<td>3 serious</td>
<td>24</td>
<td>9</td>
<td>33</td>
</tr>
<tr>
<td>4 thru 5 serious</td>
<td>15</td>
<td>16</td>
<td>31</td>
</tr>
<tr>
<td>6 or more serious</td>
<td>5</td>
<td>20</td>
<td>25</td>
</tr>
<tr>
<td>Total</td>
<td>69</td>
<td>48</td>
<td>117</td>
</tr>
</tbody>
</table>

NOTE: Chi-square=30.344, df=3, p <.001.

It is interesting to note that the Crown was almost 3 times more likely to contest bail when an accused youth had 3 factors of legal seriousness versus those youth that had 0 through 2 factors. Just over half (51.6%) of the youth who had 4 or 5 factors of legal seriousness had their bail contested, and the majority (80%) of youth with 6 or more factors also had their bail contested.

Legal Factors:

Principal Charge

Various individual components of the legal index were examined in relation to the Crown decision. Some of these factors were significant while others were not. For example, these data reveal that the principal charge before the court was not related to the Crown's decision to contest bail (Table 4).
Chi-Square = 5.63, df=3, not significant.
(Violence, drugs, b&c) pooled vs. (prop, you, ‘other cc’), not significant.

It must be pointed out however, that the principal charge is not an adequate proxy for understanding the ‘seriousness’ of an offence. There is a great deal of variation within each of these categories in terms of the actual nature of the offence. For example, within the category of violence is the charge of ‘assault’—which can take the form of a shoving match—or at the other end—an assault with a weapon. Thus the lack of a relationship here between the Crown decision and the charge category may simply be due to the actual nature of the offence—not the category it falls within.

**Prior Record**

While it appears that information about the current charges before the court did not influence the Crown decision, various kinds of legal information about the youth’s past, such as prior record, did seem to be related to the Crown’s decision to contest bail.
Table 5 shows the Crown's decision to contest bail in relation to whether the accused youth had a prior record.

Table 5

Relationship between the accused youth having a prior record and the Crown contesting release

<table>
<thead>
<tr>
<th>has Crown contest release?</th>
<th>no</th>
<th>yes</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>is there a prior record?</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>no</td>
<td>22</td>
<td>18</td>
<td>40</td>
</tr>
<tr>
<td></td>
<td>55.0%</td>
<td>45.0%</td>
<td>100.0%</td>
</tr>
<tr>
<td>yes</td>
<td>13</td>
<td>29</td>
<td>42</td>
</tr>
<tr>
<td></td>
<td>31.0%</td>
<td>69.0%</td>
<td>100.0%</td>
</tr>
<tr>
<td>not mentioned</td>
<td>30</td>
<td>1</td>
<td>31</td>
</tr>
<tr>
<td></td>
<td>96.8%</td>
<td>3.2%</td>
<td>100.0%</td>
</tr>
<tr>
<td>Total</td>
<td>65</td>
<td>48</td>
<td>113</td>
</tr>
<tr>
<td></td>
<td>57.5%</td>
<td>42.5%</td>
<td>100.0%</td>
</tr>
</tbody>
</table>

NOTE: Fisher Exact test (2-tail) p = .044 (no vs. yes); Chi-square = 31.786, df = 2, p < .001 (overall - incl not mentioned’s)\(^{13}\)

If an accused youth did not have a prior record, the Crown contested release in less than half (45%) of the cases. However, for those accused youths who did have a record, more than two-thirds (69%) had their bail contested by the Crown. Having a prior record then, seems to be one of the legal variables which influences the Crown's decision to contest bail, and as already seen in Table 2, the Crown's decision has a high degree of influence on the overall outcome of the case.

\(^{13}\) In certain cases (see Tables 5 & 7) there were instances when information on the case was 'not mentioned' in court. These were generally matters where the Crown did not contest bail for the youth. For prior record, in 30 out of 31 cases (97%) where information was 'not mentioned', the Crown consented to release the youth. This was also true for the youth having a history of bail abuse, 55/73 cases (75.3%), and previous pre-trial detention, 57/81 cases (70.4%). In all likelihood, the cases where information was 'not mentioned' are cases where the Crown has decided not to 'show cause' for detention, and in the interests of efficiency, not all details are read out loud about the case in court.
Looking more closely at the effects of prior record reveals that youth who had a prior record, which included offences that were related to the current charge before the court, were more likely to have their cases contested by the Crown (Table 6).

Table 6 - Percentage of cases contested by the Crown as a function of the relationship between prior record and current charges

<table>
<thead>
<tr>
<th>Relationship between the current charge and the nature of the prior record</th>
<th></th>
<th></th>
<th></th>
<th>sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>are priors related?</td>
<td>unstated</td>
<td>unrelated</td>
<td>related</td>
<td></td>
</tr>
<tr>
<td></td>
<td>37.5%(8)</td>
<td>50.0%(12)</td>
<td>90.9%(22)</td>
<td><em>p=.005</em></td>
</tr>
<tr>
<td>are priors recent?</td>
<td>unstated</td>
<td>not recent (over 1 year)</td>
<td>recent (within 1 year)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>44.4%(9)</td>
<td>62.5%(8)</td>
<td>80%(25)</td>
<td>n.s.</td>
</tr>
<tr>
<td>do priors include violence?</td>
<td>unstated</td>
<td>does not include violence</td>
<td>does include violence</td>
<td></td>
</tr>
<tr>
<td></td>
<td>71.4%(14)</td>
<td>58.3%(12)</td>
<td>75%(16)</td>
<td>n.s.</td>
</tr>
</tbody>
</table>

* Fisher’s exact test (2-tailed) related vs. unrelated *p=.013

A youth with unrelated prior charges had a 50-50 chance of having their bail contested, but when it was mentioned that the youth’s prior offences were related to the current allegation before the court, a very high proportion (90.9%) of these cases were contested by the Crown. This may result from the Crown perceiving the youth to be more likely to engage in a pattern of offending behaviour specific to the allegation(s) before the court, which may, in the Crown’s view, heighten the possibility that the accused youth would not remain crime free before the trial date.

Thus, prior charges that were related to the current offence may have been seen as being related to the secondary ground - the protection of the public. However, whether

---

14 This title (and the titles of subsequent tables in this thesis) does not intend to imply causality.
the prior offences were recent (within the last year), or included violence did not relate to
the Crown's decision (Table 6).

Bail Abuse & Previous Pre-trial Detention

In most of the cases observed, (73/102 or 72%), information on the youth's
history of bail abuse was not mentioned in the courtroom. As can be seen in Table 7, the
Crown contested release in almost 90% of the cases (26/29) where the issue of bail abuse
was raised, indicating that the Crown only presents this information in court when there is
an intention to contest bail release.

Table 7 - Percentage of cases contested by the Crown as a function of information
on bail abuse or pre-trial detention

<table>
<thead>
<tr>
<th></th>
<th>Yes</th>
<th>No</th>
<th>not mentioned in court</th>
<th>sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>does accused have history of bail abuse?</td>
<td>100%(10)</td>
<td>84.2%(19)</td>
<td>24.7%(73)</td>
<td>-no vs. yes - n.s.  -overall including not mentioned's p&lt;.001*</td>
</tr>
<tr>
<td>Has youth previously been in pre-trial detention?</td>
<td>86.7%(15)</td>
<td>100%(7)</td>
<td>29.6%(81)</td>
<td>-no vs. yes - n.s.  -overall including not mentioned's p&lt;.001**</td>
</tr>
</tbody>
</table>

NOTE: If the minimum expected count is less than 5, this information will be noted from here on. The
lowest expected value is indicated in parentheses below.
*1 expected value less than 5(4.31)
**2 expected values less than 5 (2.99)

The same can be seen in relation to previous pre-trial detention where the Crown
contested over 90% of the cases (20/22) when this information was mentioned.

Reverse Onus

Having a reverse onus on a case means that the accused must "show cause" why
he or she should not be detained pending trial, whereas normally the onus is on the Crown
to show cause for detention. A reverse onus condition is created in four kinds of
circumstances as set out in section 515(6) of the *Criminal Code* and applies to both youth and adults. These are:

- where the accused person is charged with committing an indictable offence while awaiting trial on another indictable offence,
- if the accused is charged with non-compliance with the conditions of her/his previous release while awaiting trial on any other charge.
- if the accused is charged with an indictable offence and is not normally a resident of Canada
- if the accused is charged with certain offences under the *Controlled Drugs and Substances Act*. 

These data show that most of the youths (78/112 or almost 70%) in this sample had to 'show cause' why they should not be detained, and as Table 8 shows, having a reverse onus was not significantly related to the Crown’s decision to contest bail release (using the traditional "0.05" standard of statistical significance).

**Table 8**

<table>
<thead>
<tr>
<th>Who has the onus?</th>
<th>does Crown contest release?</th>
<th></th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>reverse</td>
<td>no</td>
<td>39</td>
<td>50.0%</td>
</tr>
<tr>
<td></td>
<td>yes</td>
<td>39</td>
<td>50.0%</td>
</tr>
<tr>
<td>Crown onus</td>
<td>no</td>
<td>25</td>
<td>73.5%</td>
</tr>
<tr>
<td>or not mentioned</td>
<td>yes</td>
<td>9</td>
<td>26.5%</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>64</td>
<td>57.1%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>48</td>
<td>42.9%</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>100.0%</td>
</tr>
</tbody>
</table>

NOTE: Fisher's exact test (2-tail) $p = .056$

While only approaching the level of significance, the trend that appears to emerge from this table is that when there was a clear reverse onus situation the Crown appeared to be
equally likely to contest or consent to bail, but in other circumstances, such as when the
onus was on the Crown, or when this information was not mentioned in court, the Crown
seemed to be more likely to consent to the release of the youth in this sample of cases
(73.5%). It appears then, that to an extent, the reverse onus condition may serve to ‘flag’
the case to the Crown as one deserving attention and, often, detention.

In cases where it was mentioned, the most common reason for having a reverse
onus was that the youth was charged with committing an indictable offence while awaiting
trial on another indictable offence (Table 9).¹⁶

| Relationship between the reason for reverse onus (when mentioned) and the  |
| Crown decision to contest release                                      |
|---------------------------------------------------------------|------------------|--------|
| does Crown contest release?                                    | no    | yes | Total |
| if reverse onus, why?                                         |       |     |       |
| commit offence while awaiting trial                           | 22    | 24  | 46    |
| breach probation                                              | 5     | 6   | 11    |
| offence under Controlled Drugs & Substances Act               | 5     | 1   | 6     |
| Total                                                         | 32    | 31  | 63    |

NOTE: Chi-square=2.829, df=2, not significant.

Given the large effect that prior record had on the Crown’s decision to contest
release (Table 2), it is important to explore the relationship between the onus in a case and

¹⁵ for example, simple possession does not warrant a reverse onus, but trafficking constitutes a reverse onus situation.
¹⁶ Any violation of a condition of release is considered to be an indictable offence. Previous research suggests that breaches of release conditions account for the majority of reverse onus cases (Gandy 1992). In this study though, the specific offences leading up to the indictable offence charge were not specified in court, so all of this was captured under the category of ‘indictable offences’.
the Crown's decision, controlling for prior record. Table 10 shows that when the accused youth did not have a prior record, the Crown contested release in a higher proportion of cases when it was a reverse onus condition.

Table 10 - Percentage of cases contested by the Crown as a function of prior record by onus

<table>
<thead>
<tr>
<th>Prior Record</th>
<th>Reverse Onus</th>
<th>Crown Onus or Not Mentioned</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prior Record</td>
<td>72.4% (29)</td>
<td>66.7% (12)</td>
<td>n.s.</td>
</tr>
<tr>
<td>No Prior Record</td>
<td>63% (27)</td>
<td>10% (10)</td>
<td>p=.008*</td>
</tr>
</tbody>
</table>

*Fisher's exact test

However, there was not a significant relationship between the Crown decision to contest and the onus in the case when the accused youth did have a prior record.

It appears that either having the case identified legally as a reverse onus case or identifying the case as one where the accused has a record, appears to flag the case as one worthy of being contested. Having both characteristics does not appear to increase the likelihood of its being contested over having just one.

Co-accused in the present offence

Looking at the cases where information about a co-accused(s) was mentioned, shows that in these cases the Crown was equally likely to contest, or consent to bail release. However, if the youth had a co-accused and the co-accused was present in court, this was related to the Crown decision to contest bail (Table 11).
Table 11 - Percentage of cases contested by the Crown as a function of information about a co-accused(s) in the case

<table>
<thead>
<tr>
<th>does one+ charges include one+ co-accused(s)?</th>
<th>yes (%)</th>
<th>no (%)</th>
<th>not mentioned (%)</th>
<th>sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>46.3%(54)</td>
<td>52.5%(40)</td>
<td>9.5%(21)</td>
<td>-no vs. yes - n.s.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>are one or more co-accused(s) present?</th>
<th>yes (%)</th>
<th>no (%)</th>
<th>not mentioned (%)</th>
<th>sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>46.3%(54)</td>
<td>52.5%(40)</td>
<td>9.5%(21)</td>
<td>-overall including not mentioned’s p=.003</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>is one+ co-accused(s) an adult?</th>
<th>yes (%)</th>
<th>no (%)</th>
<th>not mentioned (%)</th>
<th>sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>50%(8)</td>
<td>40.6%(32)</td>
<td>50%(10)</td>
<td>n.s.</td>
</tr>
</tbody>
</table>

*Fisher’s exact test

If a co-accused was in court for the bail hearing, the Crown was less likely to contest release than if a co-accused was not present. The Crown may see the likelihood of the accused committing further offences as less probable if a co-accused is either in detention, and thus incapacitated, or ‘responsible’ enough to be present in the body of the court.

This relates, then, to both the primary and secondary grounds for detention-- that the accused is responsible enough to show up for the trial, and that further offences are unlikely to be committed if the accused is released since the co-accused is present.

Identifying a co-accused as an adult however was not significantly related to the Crown decision.

Discussion of Legal Factors:

We know that the Crown’s interpretation of a case is very important in the overall detention decision. As suggested in this section, legal variables seem to influence the Crown’s decision -- specifically, having a prior record, and having a prior record which includes offences related to the current offence before the court. Having a co-accused present for the hearing makes the Crown less likely to contest the release of the accused youth.
‘Child-like’ vs. ‘Adult-like’ Index

I created another index which included variables such as the youth’s age as well as their perceived age\(^7\), if they regularly attended school or worked full-time, whether they lived with parent(s) or if they lived on their own, among other factors (see Appendix C). My hypothesis was that the Crown might be affected by a youth who ‘looked’ or had circumstances that made them more ‘child-like’, and vice-versa. The data showed that there was no such relationship — as there were no significant differences in the Crown decision between youth who exhibited ‘adult-like qualities’ or ‘child-like qualities’ and the rest of the youth in the sample. Again, this may be due to the fact that Crown decision-making is occurring outside of the actual bail hearing.

Table 12

<table>
<thead>
<tr>
<th>How many factors are associated with being a child or adult?</th>
<th>does Crown contest release?</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>no</td>
<td>yes</td>
</tr>
<tr>
<td>adult or no child-like</td>
<td>10</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>66.7%</td>
<td>33.3%</td>
</tr>
<tr>
<td>1 child-like</td>
<td>17</td>
<td>13</td>
</tr>
<tr>
<td></td>
<td>56.7%</td>
<td>43.3%</td>
</tr>
<tr>
<td>2 child-like</td>
<td>27</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td>69.2%</td>
<td>30.8%</td>
</tr>
<tr>
<td>3 or more child-like</td>
<td>15</td>
<td>18</td>
</tr>
<tr>
<td></td>
<td>45.5%</td>
<td>54.5%</td>
</tr>
<tr>
<td>Total</td>
<td>69</td>
<td>48</td>
</tr>
<tr>
<td></td>
<td>56.0%</td>
<td>41.0%</td>
</tr>
</tbody>
</table>

NOTE: Chi-square = 4.622, \(df = 3\), not significant.

Analyses on the individual factors that are contained in this index also showed no significant effects.
Demographic Factors in the Bail Decision

Age

There has been very little research on the effect of age as a factor in decision-making in youth courts, and what research has been done shows mixed results (Wordes, Bynum & Corley 1994, Carrington, Moyer and Kopelman 1988, Bookin-Weiner 1984, Frazier & Bishop 1985). As mentioned before, there was an expectation that the courts would be inclined towards treating ‘younger’ youths less harshly than their older counterparts (Kowalski & Caputo 1999). If this were to be the case, it would be consistent with recent research in developmental psychology which has pointed out that younger teens differ substantially from older teens and adults in their cognitive capacity, attitudes and perceptions of risk (Scott & Grisso 1997).

Looking at the question of age in relation to Crown decision-making for bail hearings shows that there were no differences in the decision of the Crown based upon which age group the accused youth was in (12-13, 14-15 or 16 & older) (Table 13).

Table 13

<table>
<thead>
<tr>
<th>Age Group</th>
<th>Does Crown Contest Release?</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>12-13</td>
<td>6</td>
<td>5</td>
</tr>
<tr>
<td>14-15</td>
<td>20</td>
<td>14</td>
</tr>
<tr>
<td>16+</td>
<td>37</td>
<td>29</td>
</tr>
<tr>
<td>Total</td>
<td>63</td>
<td>48</td>
</tr>
</tbody>
</table>

NOTE: Chi-Square = .094, df = 2, not significant

17 As assessed by the researchers in court
A number of other demographic variables, and factors relating to how the youth ‘appeared’ in court were recorded. The results of this analysis show that none of these details about the youth were related to the Crown decision to contest bail (Table 14).

Table 14 - Percentage of cases where the Crown contested bail release as a function of demographic and appearance-related factors

<table>
<thead>
<tr>
<th></th>
<th>male</th>
<th>female</th>
<th>sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>gender</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>male</td>
<td>female</td>
<td></td>
</tr>
<tr>
<td></td>
<td>43.8%(96)</td>
<td>28.6%(21)</td>
<td>n.s.</td>
</tr>
<tr>
<td><strong>perceived ethnicity?</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>black</td>
<td>51.3%(39)</td>
<td>35.1%(27)</td>
<td>n.s.</td>
</tr>
<tr>
<td>other*</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>are there ‘adult-like’ signs?</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>yes</td>
<td>60.0%(5)</td>
<td>40.2%(112)</td>
<td>n.s.</td>
</tr>
<tr>
<td>no</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>are there ‘child-like’ signs?</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>no</td>
<td>40.4%(109)</td>
<td>50.0%(8)</td>
<td>n.s.</td>
</tr>
<tr>
<td>yes</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>how is youth dressed for court?</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Not dressed up</td>
<td>42.0%(100)</td>
<td>35.3%(17)</td>
<td>n.s.</td>
</tr>
<tr>
<td>dressed up</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>does youth have facial hair? (males only)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>yes</td>
<td>38.1%(21)</td>
<td>42.7%(82)</td>
<td>n.s.</td>
</tr>
<tr>
<td>no</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>is youth interested in proceedings?</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>not interested</td>
<td>56.7%(39)</td>
<td>35.7%(84)</td>
<td>p=.054**</td>
</tr>
<tr>
<td>very/moderately interested</td>
<td></td>
<td></td>
<td>(n.s.)</td>
</tr>
</tbody>
</table>

* the category ‘other’ includes, caucasian, south asian, aboriginal, hispanic or other.
**Fisher’s exact test

Gender:

There was a very small proportion of accused female youths in this sample. These data do not reveal significant differences in the Crown decision to contest bail based upon the gender of the accused. And in addition these results show that the perceived ethnicity
of the accused youth by the researchers was not related to the Crown decision to detain (Table 14).

**How the youth ‘appeared in court’**

Factors relating to how the youth appeared in court,\(^\text{18}\) which were subjectively assessed by the researchers, were not related to the Crown decision (Table 14). Obviously, one explanation for this is that the majority of decisions made by the Crown are formulated outside of the bail hearing. Alternatively, they may simply be irrelevant to the Crown in deciding whether to contest release, or, the youth in these bail hearings did not stand out remarkably from one another in terms of how they looked, how interested they were in court, or if they exhibited ‘child-like’ or ‘adult-like’ qualities. Generally, these youths were quiet in court, and the majority of the accused youths (72%) were observed by the researchers to be ‘very’ or ‘moderately’ interested in the court proceedings. As well, most of the youth in court (85.6%) were perceived by the researchers to be ‘dressed-down’ in court — meaning they were wearing clothes that teenagers generally wear — jeans, t-shirts, running shoes — as opposed to the remainder (14.4%) that were dressed in a ‘neat-middle-class’ way or were ‘dressed-up’ for court.

The fact that most of the youth observed in the bail hearings were not ‘dressed up’ for court may be accounted for by the fact that most of these youth (89.8%) were detained in secure detention after being arrested. The only way for an accused person to change their clothes for court from what they were wearing when arrested would be if a family

---

\(^{18}\) such as how was the youth dressed for court; was the youth interested in the proceedings; how old did the youth look to the researchers, etc.
member or other supportive person was able to bring a change of clothes to them at their place of detention.

Adult or Child-like signs:

As also seen in Table 14, in only 5 cases did we find something noteworthy about the youth as being ‘adult-like’, and in 3 of these 5 cases, the Crown contested release. Similarly, in only 8 cases did we notice that the accused before the court exhibited characteristics that might be perceived to be ‘child-like’. Here, the Crown contested half of these cases.

For the males in these cases, having facial hair was quite rare. In only 21 cases (18%) did the youth have any kind of facial hair. Of these cases, 8 (or 38%) were contested by the Crown. Again, the hypothesis was that more ‘adult-like’ characteristics (having facial hair) might be related to the decision to detain. Clearly this was not the case.

Discussion of Demographic and ‘Child-like’ vs. ‘Adult-like’ factors:

The youth’s appearance in court had little to do with the decisions the Crown was making. Thus, the Crown’s approach to a case was related more to factors surrounding the youth’s legal history than present appearances in the bail hearing. In addition, these data reveal that for the most part the accused youths in this sample were a remarkably homogeneous group, and thus, even if the Crown was ‘looking’ at these youths in terms of decision-making, there was little variation which would have made particular youths stand apart.
Support Index

In trying to determine how much of a risk an accused youth might pose if released on bail, the Crown may be likely to assess the youth’s network of support. Such factors as having a parent or guardian present at the court hearing, having knowledge that the parent is involved in the youth’s life and is able to supervise, and having a defense lawyer who provides a case plan for the youth upon release, all speak to the degree to which others in the youth’s life are able and willing to provide support. The data in this sample show that there was no significant relationship between the amount of support they had and the Crown decision to contest release (Table 15).

Table 15

<table>
<thead>
<tr>
<th>How many factors are associated with support for this youth?</th>
<th>does Crown contest release?</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>no</td>
<td>yes</td>
<td>Total</td>
</tr>
<tr>
<td>0 support</td>
<td>18</td>
<td>7</td>
<td>25</td>
</tr>
<tr>
<td>1-2 support</td>
<td>72.0%</td>
<td>28.0%</td>
<td>100.0%</td>
</tr>
<tr>
<td>3 or more support</td>
<td>21</td>
<td>18</td>
<td>39</td>
</tr>
<tr>
<td></td>
<td>53.8%</td>
<td>46.2%</td>
<td>100.0%</td>
</tr>
<tr>
<td></td>
<td>30</td>
<td>23</td>
<td>53</td>
</tr>
<tr>
<td></td>
<td>55.6%</td>
<td>43.4%</td>
<td>100.0%</td>
</tr>
<tr>
<td>Total</td>
<td>69</td>
<td>48</td>
<td>117</td>
</tr>
<tr>
<td></td>
<td>59.0%</td>
<td>41.0%</td>
<td>100.0%</td>
</tr>
</tbody>
</table>

NOTE: Chi-square = 2.300, df = 2, not significant.
Family & School

In the cases observed in this study, at least one parent or parent figure was present for the bail hearing the majority of the time (72%). However, having a parent present in court was not significantly related to the Crown decision to contest release (Table 16).

Table 16 - Percentage of cases where Crown contests bail as a function of parental presence in court, involvement in youths’ life and supervision

<table>
<thead>
<tr>
<th>parent(s) present?</th>
<th>yes</th>
<th>no</th>
<th>sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>does court hear that parent(s) is involved in youth's life?</td>
<td>43.9%(82)</td>
<td>37.5%(32)</td>
<td>n.s.</td>
</tr>
<tr>
<td>does court hear that parent(s) able to supervise youth?</td>
<td>44.3%(79)</td>
<td>72.7%(11)</td>
<td>n.s.</td>
</tr>
<tr>
<td></td>
<td>42.9%(56)</td>
<td>70.0%(20)</td>
<td>p=.067* n.s.</td>
</tr>
</tbody>
</table>

* Fisher's exact test

Living Arrangements

However, living at home was a significant factor in the Crown's decision to contest bail. As Table 17 shows, over two-thirds (69.6%) of youths who did not live at home had their bail contested by the Crown, while a smaller proportion (37.5%) of youth who lived at home had their bail contested. This is consistent with previous research which has shown that youth who were detained were less likely to reside with two parents (Wordes, Bynum & Corley 1994, Schutt & Dannefer 1988).
Table 17

Relationship between the youth living with their parent(s)/guardian(s) and the Crown decision to contest release

<table>
<thead>
<tr>
<th>Does the youth live with their parent(s)/guardian(s)?</th>
<th>does Crown contest release?</th>
<th></th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>no</td>
<td>7</td>
<td>16</td>
</tr>
<tr>
<td></td>
<td>yes</td>
<td>30.4%</td>
<td>68.6%</td>
</tr>
<tr>
<td>yes</td>
<td>50</td>
<td>30</td>
<td>80</td>
</tr>
<tr>
<td></td>
<td>62.5%</td>
<td>37.5%</td>
<td>100.0%</td>
</tr>
<tr>
<td>not mentioned</td>
<td>11</td>
<td>2</td>
<td>13</td>
</tr>
<tr>
<td></td>
<td>84.6%</td>
<td>15.4%</td>
<td>100.0%</td>
</tr>
<tr>
<td>Total</td>
<td>68</td>
<td>48</td>
<td>116</td>
</tr>
<tr>
<td></td>
<td>58.6%</td>
<td>41.4%</td>
<td>100.0%</td>
</tr>
</tbody>
</table>

NOTE: Fisher exact test (2-tail) $p = .009$ (no vs. yes).
Chi-square = 11.651, $df = 2$, $p = .003$ (overall including not mentioned’s).

Clearly, whether or not an accused youth lives with a parent is highly related to the
Crown’s decision on whether or not to contest bail. It directly relates to the primary
ground for bail detention, which is to secure the youth’s attendance for trial. Thus, an
accused youth who lives with a parent would be seen to have support sufficient to ensure
attendance at trial. However, this finding raises issues relating to the treatment of accused
youths who do not have the ability to live at home for various reasons—and the possibility
that detention is being used more frequently for youth who do not live with their parents.
The fundamental problem for researchers on this issue is that the primary ground for
detention—the likelihood of attending trial—directs the court’s attention to the home
situation of the young person (Moyer 1996). But, the use of detention for child welfare
reasons needs to be disentangled from the use of detention to ensure that the youth shows
up for trial—otherwise detention is unjustified (Nasmith 1990, Gandy 1992, Guggenheim
1977).
School attendance

School attendance was also related to the Crown’s decision in the bail hearings observed.\(^{19}\) Table 18 shows that the Crown contested all of the cases where the youth was not in school, whereas if a youth attended school regularly, just over two-thirds (69.4\%) had their case contested.

### Table 18

<table>
<thead>
<tr>
<th>does Crown contest release?</th>
<th>no</th>
<th>yes</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Is the youth in school?</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>no</td>
<td>00</td>
<td>16</td>
<td>16</td>
</tr>
<tr>
<td>yes</td>
<td>11</td>
<td>25</td>
<td>36</td>
</tr>
<tr>
<td>Total</td>
<td>18</td>
<td>41</td>
<td>59</td>
</tr>
<tr>
<td>Is the youth in school?</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>yes</td>
<td>11</td>
<td>25</td>
<td>36</td>
</tr>
<tr>
<td>Total</td>
<td>18</td>
<td>41</td>
<td>59</td>
</tr>
<tr>
<td>not mentioned</td>
<td>57</td>
<td>7</td>
<td>64</td>
</tr>
<tr>
<td>Total</td>
<td>58</td>
<td>48</td>
<td>106</td>
</tr>
</tbody>
</table>

NOTE: Fisher’s exact test (2-tail) \(p = .012\) (no vs. yes).
Chi-square = 58.807, \(df=2\), \(p<.001\) (overall - incl. not mentioned’s).

School attendance, however, was also related to another factor which appears to relate to the Crown decision – prior record (see Table 5). As Table 19 reveals, there was a substantial relationship between not attending school regularly, and having a prior record. A very high proportion (87.5\%) of those youth who were not attending school, had a prior record, whereas, the majority (60\%) of youths who were attending school regularly, did not have a prior record.

\(^{19}\) The variable for the youth being in school was pooled where ‘yes’ included that the youth was in school ‘most of the time’, in ‘summer school’ or ‘yes, was in school’. And ‘no’ included ‘no, not in school’, in school ‘once in a while’, ‘not registered’, ‘kicked out’ or ‘suspended’.
Fisher's Exact test (2 - tail) \( p = .002 \).

However, the relationship between the youth being in school and the Crown decision to contest release was not significant when controlling for prior record (Table 20). For both groups -- those youth with and those without a prior record -- the relationship was not significant, in large part because of the lack of variation in school attendance for those without a record and in the Crown's decision for those with a record.

**Table 20** - Percentage of cases contested by the Crown as a function of prior record by youth attending school

<table>
<thead>
<tr>
<th>prior record</th>
<th>youth attends school</th>
<th>youth does not attend school</th>
<th>sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>prior record</td>
<td>91.7%(12)</td>
<td>100%(14)</td>
<td>n.s.</td>
</tr>
<tr>
<td>no prior record</td>
<td>77.8%(18)</td>
<td>100%(2)</td>
<td>n.s.</td>
</tr>
</tbody>
</table>
Discussion of Support Factors:

In trying to understand how factors of support relate to the Crown decision, these data seem to suggest that exhibiting 'ties to the community' in the form of attending school, and 'ties to the family' by living with one's parents are factors that are relevant for the Crown in determining the risk at youth poses if released. Both address the amount of supervision a youth has which may mean there is less opportunity to engage in further offences pending trial. Interestingly, on the issue of support, having a parent present in court did not impact on the Crown decision—again possibly owing to the fact that the Crown may be making the decision outside of the bail hearing.

**Which factors overall predict the Crown decision on bail release?**

To conclude this section, it is important to understand which (if any) of these factors may provide more predictive ability for the Crown decision than the others. Table 21 provides an ordinary least squares regression\(^{20}\) of the significant factors that related to the Crown's decision.

---

\(^{20}\) A logistic regression analysis might have been seen as being the most appropriate test for the predictability of these variables since the dependent variable (does the Crown contest?) is binary. However, the sample size is too small for logistic regression and thus, ordinary least squares regression analysis was performed in order to analyze estimated effects of the independent variables. This is also the case for all regression analyses in this chapter.
Table 21

Ordinary Least Squares Regression analysis representing estimated effects of independent variables on the Crown’s decision to contest release

<table>
<thead>
<tr>
<th>Independent Variables</th>
<th>unstandardized coefficients</th>
<th>standardized coefficients</th>
<th>t</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Constant)</td>
<td>2.170</td>
<td>.351</td>
<td>6.186</td>
<td>.000</td>
</tr>
<tr>
<td>prior record</td>
<td>.263</td>
<td>.093</td>
<td>2.812</td>
<td>.006</td>
</tr>
<tr>
<td>does youth live with parent(s)?</td>
<td>-.134</td>
<td>.105</td>
<td>-1.273</td>
<td>.205</td>
</tr>
<tr>
<td>in school?</td>
<td>-.462</td>
<td>.130</td>
<td>-3.553</td>
<td>.001</td>
</tr>
</tbody>
</table>

*a Dependent Variable: does Crown contest release?

NOTE: dependent variable: ‘Crown contest’ (1=no 2=yes)
independent variables: ‘prior record’ (1=no / not mentioned 2=yes)
‘live with parent’ (1=no 2=yes or not mentioned)
‘in school?’ (1=no 2= yes or not mentioned)

Of all the factors which were significantly related to the Crown’s decision, both having a prior record and the youth being in school have independent effects on the likelihood of the Crown contesting release. If there was a prior record, the Crown was more likely to contest bail release, and if the youth was said not to be in school, the Crown was also more likely to contest bail. However, whether a youth lived with parent(s) or not did not have independent effects on the predictability of the Crown’s decision.

Conclusions: The decision-making process of the Crown:

Clearly, the Crown decision to contest bail in youth court is highly influential in the overall bail decision (Table 2). It appears from these data then, that the Crown decision is based on legal factors relating to the youth’s history in the criminal justice system - specifically - prior record. However, also related to the Crown decision are variables which speak to the ‘ties’ to the community -- does the youth live with parent(s)/or guardian(s)? Is the youth attending school regularly? Although, as the regression analysis
in Table 21 reveals, only the variable of ‘school attendance’ adds significantly to the predictability of the Crown decision. It could be argued that rather than being strictly looked at as a tie to the community, school attendance is a variable which relates to the youthfulness of the offender -- issues of supervision are more important for the younger offender. This piece of information about the youth’s circumstances would also be known to the Crown prior to the hearing. The other in-court factors that were tested, however, (e.g. youth’s demeanour in court) are not related to the Crown decision. One possible explanation is that Crown decision-making occurs outside of the bail hearing, and since legal factors are of primary importance, the Crown may not be able to even assess these in-court factors.

Thus, in bail decision-making for youth, the first critical point seems to lie with the Crown attorney. When the Crown consents to release or opts not to ‘show cause’, all accused youths are released, but in the cases where the Crown contests release, it is up to the justice or judge to determine whether to grant bail or not. It is an examination of the factors that the justice or judge is responding to that we will turn to next.

Section II - Factors affecting the decision to grant bail when the Crown contests release

Legal Index:

There were 48 cases out of 118 (40.6%) where the Crown contested the release of the youth. Because the sample size is so small, many of the analyses should be considered to be suggestive only. An effect would have to be quite large (e.g. seriousness of the charge) in order to be statistically significant (or even appropriate for the use of inferential statistics). This is due in part to the fact that some of the events being observed for this
study occurred quite infrequently. As such, this section examines these relationships, but as will become apparent, many of these relationships are not statistically significant.

An analysis of the index of legal factors shows that there was not a significant relationship between legal seriousness and the decision to grant bail for these contested cases (Table 22).

**Table 22**

<table>
<thead>
<tr>
<th>Factors of legal seriousness</th>
<th>no</th>
<th>yes</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 through 2 serious</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>3 serious</td>
<td>5</td>
<td>4</td>
<td>9</td>
</tr>
<tr>
<td>4 thru 5 serious</td>
<td>12</td>
<td>4</td>
<td>16</td>
</tr>
<tr>
<td>6 or more serious</td>
<td>17</td>
<td>3</td>
<td>20</td>
</tr>
<tr>
<td>Total</td>
<td>35</td>
<td>13</td>
<td>48</td>
</tr>
</tbody>
</table>

Chi-square=5.268, df=3, not significant.

However, while the relationship was not statistically significant, it appears that the tendency is still in the same direction as seen in Section I with the Crown. With more factors of legal seriousness, there appeared to be a greater likelihood that an accused youth would be denied bail.
Legal Factors

Principal Charge

Table 23 shows the relationship between categories of the principal charge and the decision to grant bail for the 48 contested cases. This table suggests that there was a relationship between the allegations before the court in the cases that the Crown was contesting and the outcome. It must again be noted however, that the principal charge is not an adequate proxy for the actual seriousness of the charge. However, by grouping the category of these offences in order to get some insight into how charges may affect decision-making shows that bail was more likely to be denied in cases that included violence, drugs or break & enter even though at the time of the bail hearing these are allegations before the court. The large majority (82.9%) of cases, which included potentially more serious kinds of offences, were not granted bail.

Table 23

<table>
<thead>
<tr>
<th>principal charge</th>
<th>was bail granted?</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>no</td>
<td>yes</td>
<td>Total</td>
<td></td>
</tr>
<tr>
<td>violence</td>
<td>19</td>
<td>3</td>
<td>22</td>
<td></td>
</tr>
<tr>
<td></td>
<td>86.4%</td>
<td>13.6%</td>
<td>100.0%</td>
<td></td>
</tr>
<tr>
<td>drugs</td>
<td>4</td>
<td>1</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td></td>
<td>80.0%</td>
<td>20.0%</td>
<td>100.0%</td>
<td></td>
</tr>
<tr>
<td>b&amp;e</td>
<td>6</td>
<td>2</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td></td>
<td>75.0%</td>
<td>25.0%</td>
<td>100.0%</td>
<td></td>
</tr>
<tr>
<td>prop&amp;yco&amp;cc</td>
<td>6</td>
<td>7</td>
<td>13</td>
<td></td>
</tr>
<tr>
<td></td>
<td>46.2%</td>
<td>53.8%</td>
<td>100.0%</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>35</td>
<td>13</td>
<td>48</td>
<td></td>
</tr>
<tr>
<td></td>
<td>72.9%</td>
<td>27.1%</td>
<td>100.0%</td>
<td></td>
</tr>
</tbody>
</table>

Chi-square=6.874, df=3, n.s. (overall)
Statistics for categories of charges pooled ‘violence, drugs, b&e’ vs. ‘property, yco & ‘other cc’”; Fisher’s exact test (2-tail) p=.024
One might expect that the cases with more serious charges may also be cases where the youth has prior convictions. When controlling for the effects of prior record, Table 24 shows that when a youth did not have a prior record, they were still more likely to be detained if the offence they committed was of a serious nature. The same relationship did not hold for those youth with a prior record, though it was in the same direction.

**Table 24** - Percentage of cases denied bail as a function of prior record by the category of principal charge(s) before the court for cases where the Crown contested bail

<table>
<thead>
<tr>
<th></th>
<th>'less serious' charges</th>
<th>'more serious' charges</th>
<th>sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>property, POA, other CC</td>
<td>violence, drugs, b&amp;l</td>
<td></td>
</tr>
<tr>
<td>prior record</td>
<td>62.5%(8)</td>
<td>85.7%(21)</td>
<td>n.s.</td>
</tr>
<tr>
<td>no prior record</td>
<td>20.0%(5)</td>
<td>76.9%(13)</td>
<td>(p=.047^*)</td>
</tr>
</tbody>
</table>

*Fisher's exact test

**Prior Record**

While prior record was highly influential in predicting Crown decision-making, it was not related significantly to the decision to grant bail in the cases where the Crown contested release. However the direction in Table 25 suggests that a youth with a prior record may have a higher likelihood of being denied bail than a youth with no prior record.
Table 25 - Percentage of cases where bail was denied as function of prior record information for only those cases where the Crown contested release

<table>
<thead>
<tr>
<th>does youth have a prior record?</th>
<th>unstated</th>
<th>no</th>
<th>yes</th>
<th>sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>n/a</td>
<td>61.1%(18)</td>
<td>79.3%(29)</td>
<td>n.s.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>are priors related?</th>
<th>unstated</th>
<th>unrelated</th>
<th>related</th>
<th>sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>66.7%(3)</td>
<td>66.7%(4)</td>
<td>85.0%(20)</td>
<td>n.s.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>are priors recent?</th>
<th>unstated</th>
<th>not recent (over 1 year)</th>
<th>recent (within 1 year)</th>
<th>sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>75.0%(4)</td>
<td>60.0%(5)</td>
<td>85.0%(20)</td>
<td>n.s.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>do priors include violence?</th>
<th>unstated</th>
<th>does not include violence</th>
<th>does include violence</th>
<th>sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>60.0%(10)</td>
<td>100%(7)</td>
<td>83.3%(12)</td>
<td>n.s.</td>
</tr>
</tbody>
</table>

There were also no significant relationships between information about the nature of the prior record and the overall bail decision. However, in a similar pattern to the Crown’s decision-making seen in Section I, there seems to have been a higher likelihood of having bail denied when the prior record had offences that were related to the current charge or that were recent. However, having prior offences that included violence was not in the same direction as with the Crown.

History of bail abuse & previous pre-trial detention

While not statistically significant, the trend that appears to emerge from Table 26 is that the courts were slightly more inclined to deny bail in cases where information on bail abuse or previous pre-trial detention was raised in court versus it not being raised—or where the court heard that the youth was not in either of these conditions previously. This is consistent with recommendations of the Crown – who only raised these issues in the cases where release was contested.
Table 26 - Percentage of cases denied bail as a function of information on bail abuse or pre-trial detention for only those cases where the Crown contested release

<table>
<thead>
<tr>
<th></th>
<th>Yes</th>
<th>No</th>
<th>not mentioned in court</th>
<th>sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>does accused have history of bail abuse?</td>
<td>90.0%(10)</td>
<td>68.8%(16)</td>
<td>66.7%(18)</td>
<td>n.s.</td>
</tr>
<tr>
<td>has youth previously been in pre-trial detention?</td>
<td>84.6%(13)</td>
<td>71.4%(7)</td>
<td>66.7%(240)</td>
<td>n.s.</td>
</tr>
</tbody>
</table>

Reverse Onus

The relationship between the onus in the case and the decision to grant bail for the cases where the Crown contested was not statistically significant (Table 27).

Table 27

<table>
<thead>
<tr>
<th>Who has the onus?</th>
<th>was bail granted?</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>no</td>
<td>yes</td>
</tr>
<tr>
<td>reverse</td>
<td>27</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td>69.2%</td>
<td>30.8%</td>
</tr>
<tr>
<td>Crown onus or not mentioned</td>
<td>8</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>88.9%</td>
<td>11.1%</td>
</tr>
<tr>
<td>Total</td>
<td>35</td>
<td>13</td>
</tr>
</tbody>
</table>

Chi-square (corrected)=-.609, df=1, not significant.

Co-accused in the present offence

Having a co-accused was also not a significant factor in the court’s overall decision, nor was the relationship between a co-accused being present in the courtroom and the court’s decision on bail in cases contested by the Crown. Even having an adult co-accused appeared not to impact decisions in contested cases (Table 28).
Table 28 - Percentage of cases denied bail as a function of information about a co-accused(s) in the case for only those cases where the Crown contested release

<table>
<thead>
<tr>
<th>does one+ charges include one+ co-accused(s)?</th>
<th>yes</th>
<th>no</th>
<th>not mentioned</th>
<th>sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>80.0%(25)</td>
<td>66.7%(21)</td>
<td>n/a</td>
<td>n.s.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>are one or more co-accused(s) present?</th>
<th>yes</th>
<th>no</th>
<th>not mentioned</th>
<th>sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>90.0%(10)</td>
<td>76.9%(13)</td>
<td>n/a</td>
<td>n.s.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>is one+ co-accused(s) an adult?</th>
<th>yes</th>
<th>no</th>
<th>not mentioned</th>
<th>sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>100%(4)</td>
<td>84.6%(13)</td>
<td>40.0%(5)</td>
<td>n.s.</td>
</tr>
</tbody>
</table>

'Child-like' vs. 'Adult-like' Index

There was no relationship between the index of 'child-like' or 'adult-like' factors and the court's decision on bail for the cases where the Crown had already contested release perhaps because these factors occurred too infrequently.

Table 29 - Relationship between youth having 'adult' or 'child-like' qualities and the decision to grant bail for only those cases where the Crown contested bail

<table>
<thead>
<tr>
<th>How many factors are associated with being a child or adult?</th>
<th>was bail granted?</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>no</td>
<td>yes</td>
<td>total</td>
</tr>
<tr>
<td>adult or 0 child</td>
<td>4</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>80.0%</td>
<td>20.0%</td>
<td>100.0%</td>
</tr>
<tr>
<td>1 child-like</td>
<td>9</td>
<td>4</td>
<td>13</td>
</tr>
<tr>
<td></td>
<td>69.2%</td>
<td>30.8%</td>
<td>100.0%</td>
</tr>
<tr>
<td>2 child-like</td>
<td>8</td>
<td>4</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td>66.7%</td>
<td>33.3%</td>
<td>100.0%</td>
</tr>
<tr>
<td>3+ child-like</td>
<td>14</td>
<td>4</td>
<td>18</td>
</tr>
<tr>
<td></td>
<td>77.8%</td>
<td>22.2%</td>
<td>100.0%</td>
</tr>
<tr>
<td>Total</td>
<td>35</td>
<td>13</td>
<td>48</td>
</tr>
<tr>
<td></td>
<td>72.9%</td>
<td>27.1%</td>
<td>100.0%</td>
</tr>
</tbody>
</table>

Chi-square=.669, df=3, not significant.
Demographic Factors in the Bail Decision

Age

The courts did not appear to be influenced by the age group that the youth was in when making a determination about bail release. Similar to the Crown, each age group appeared to be equally likely to be granted or denied bail.

Table 30

Relationship between the age group of the accused youth and the decision to grant bail for only those cases where the Crown contested release

<table>
<thead>
<tr>
<th>age group</th>
<th>was bail granted?</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>no</td>
<td>yes</td>
<td>Total</td>
</tr>
<tr>
<td>12&amp;13</td>
<td>3</td>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>60.0%</td>
<td>40.0%</td>
<td>100.0%</td>
</tr>
<tr>
<td>14&amp;15</td>
<td>11</td>
<td>3</td>
<td>14</td>
</tr>
<tr>
<td></td>
<td>78.6%</td>
<td>21.4%</td>
<td>100.0%</td>
</tr>
<tr>
<td>16+</td>
<td>21</td>
<td>8</td>
<td>29</td>
</tr>
<tr>
<td></td>
<td>72.4%</td>
<td>27.6%</td>
<td>100.0%</td>
</tr>
<tr>
<td>Total</td>
<td>35</td>
<td>13</td>
<td>48</td>
</tr>
<tr>
<td></td>
<td>72.9%</td>
<td>27.1%</td>
<td>100.0%</td>
</tr>
</tbody>
</table>

NOTE: Chi-square=.653, df=2, not significant.
Other Demographic & Appearance Related Factors:

Table 31 - Percentage of cases where bail was denied as a function of demographic and appearance-related factors for only cases where the Crown contested release

<table>
<thead>
<tr>
<th>Gender</th>
<th>Male</th>
<th>Female</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>73.8%(42)</td>
<td>66.7%(6)</td>
<td>n.s.</td>
</tr>
<tr>
<td>Perceived Ethnicity?</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Black</td>
<td>75.0%(20)</td>
<td>70.4%(27)</td>
<td>n.s.</td>
</tr>
<tr>
<td>Other</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>'adult-like' signs?</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>66.7%(3)</td>
<td>73.3%(45)</td>
<td>n.s.</td>
</tr>
<tr>
<td>No</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>'child-like' signs?</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>79.5%(44)</td>
<td>0%(4)</td>
<td>p=.004*</td>
</tr>
<tr>
<td>No</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>How is youth dressed for court?</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Not dressed up</td>
<td>76.2%(42)</td>
<td>50.0%(6)</td>
<td>n.s.</td>
</tr>
<tr>
<td>Dressed up</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Does youth have facial hair? (males only)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>75.0%(8)</td>
<td>74.3%(35)</td>
<td>n.s.</td>
</tr>
<tr>
<td>No</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Is youth interested in proceedings?</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Not interested</td>
<td>76.5%(17)</td>
<td>70.0%(30)</td>
<td>n.s.</td>
</tr>
<tr>
<td>Very/moderately interested</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Fisher's exact test

With the exception of clearly exhibiting child-like characteristics in court, all of these demographic and 'appearance related' factors are statistically non significant. What seems to be the case is that if the Crown has contested release -- as in all of these cases -- the majority are denied bail by the court as well.

In terms of the effects of 'youthfulness' in court, as Table 31 suggests, in only 3 of the 48 contested cases did the youth exhibit 'adult-like' signs. Hence, no inferences can be made about possible effects. Interestingly though, there was a significant relationship
between the youth exhibiting 'child-like' signs and the court decision in the 48 contested cases. In all 4 cases where the youth exhibited 'child-like' signs, bail was granted.21

**Support Index**

There was not a significant relationship between the number of factors associated with support that a youth had and the court’s decision to grant bail but the trend was in the direction of bail being more likely granted in cases with a higher number of factors indicating support.

**Table 32**

<table>
<thead>
<tr>
<th>How many factors of support are associated with this youth?</th>
<th>0 support</th>
<th>1-2 support</th>
<th>3 or more support</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>was bail granted?</td>
<td>no</td>
<td>yes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>no</td>
<td>7</td>
<td>0.00</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td>yes</td>
<td>100.0%</td>
<td>100.0%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1-2 support</td>
<td>14</td>
<td>4</td>
<td>18</td>
<td></td>
</tr>
<tr>
<td>3 or more support</td>
<td>14</td>
<td>9</td>
<td>23</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>35</td>
<td>13</td>
<td>48</td>
<td></td>
</tr>
</tbody>
</table>

Chi-square=4.506, df=2, not sig.

---

21 The 4 child-like signs were (1) the youth looked very young to the researcher (2) the youth raised his hand in court to speak (3) the youth was crying and telling his mother he wouldn’t make her come back to court for him again (4) the youth was crying during the hearing.
Parental support

While not statistically significant, it appears that the courts were more likely to deny bail when a parent/guardian was not present in court (Table 33). This same pattern was true for the Crown decision to contest release (see Table 16).

Table 33 - Percentage of cases where bail was denied as a function of parental presence in court & supervision for only cases where the Crown contested release

<table>
<thead>
<tr>
<th>parent(s) present?</th>
<th>yes</th>
<th>no</th>
<th>sig</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>69.4%(36)</td>
<td>83.3%(12)</td>
<td>n.s.</td>
</tr>
<tr>
<td>does court hear that parent(s) is involved in youth's life?</td>
<td>68.6%(35)</td>
<td>75.0%(8)</td>
<td>n.s.</td>
</tr>
<tr>
<td>does court hear that parent(s) able to supervise youth?</td>
<td>66.6%(24)</td>
<td>78.6%(14)</td>
<td>n.s.</td>
</tr>
</tbody>
</table>

Living arrangements & School

Living with a parent was significantly related to the Crown decision to contest. As shown in Section I, the Crown was more likely to contest release if the youth was not living with their parent or guardian. While not significant in these 48 contested cases, the same kind of trend emerges as with the Crown decision. As seen in Table 34, in 14 out of 16 (87.5%) cases where the youth did not live with their parent or guardian, the justice or judge denied bail.
Table 34 - Percentage of cases where bail was denied as a function of living arrangements & school for cases where the Crown contested release

<table>
<thead>
<tr>
<th></th>
<th>yes</th>
<th>no</th>
<th>not mentioned</th>
<th>sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>does this youth live with parent(s)/guardian(s)?</td>
<td>66.7%(30)</td>
<td>87.5%(16)</td>
<td>50.0%(2)</td>
<td>n.s.</td>
</tr>
<tr>
<td>is this youth in school?</td>
<td>64.0%(25)</td>
<td>81.3%(16)</td>
<td>85.7%(7)</td>
<td>n.s.</td>
</tr>
</tbody>
</table>

Similarly, attending school was related to the decision to contest bail. The Crown contested all of the youth who were not attending school. Of those cases where the youth was not in school, most (81.3%) were denied bail. While not statistically significant, it appears that the courts may also be looking at school attendance in determining bail release decisions for the 48 cases contested by the Crown.

Discussion

With the exception of the category of the ‘seriousness’ of the charge and ‘child-like’ factors, all of the relationships in this section were not significantly related to the decision to grant bail for cases that were contested by the Crown. However, for the most part, the trends were in the same direction as the Crown in Section I. This suggests that after the Crown filters out cases by consenting to their release, the courts decide on the contested cases by looking at many of the same factors as the Crown did in determining bail. However, beyond these factors, the courts seem to be influenced by the category of the charge—even when the youth does not have a prior record.

As well, the courts may be somewhat influenced by youth that exhibit ‘child-like’ characteristics in the courtroom, perhaps responding to the youthfulness of an accused youth at the time of their hearing. However, only in these very rare situations were the courts responding to ‘child-like’ factors.
Given that ‘child-like’ factors and the principal charge are significantly related to the bail decision in contested cases, it is useful to determine which of these factors results in a higher level of predictability for the decision to grant bail after the Crown has contested.

Table 35

Ordinary Least Squares Regression analysis representing estimated effects of independent variables on decision to grant bail in Crown contested cases

<table>
<thead>
<tr>
<th>Independent Variables</th>
<th>B</th>
<th>Std. Error</th>
<th>Beta</th>
<th>t</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Constant)</td>
<td>1.004</td>
<td>.330</td>
<td></td>
<td>3.040</td>
<td>.004</td>
</tr>
<tr>
<td>principal charge</td>
<td>-.297</td>
<td>.123</td>
<td>-.297</td>
<td>-2.409</td>
<td>.020</td>
</tr>
<tr>
<td>'child-like' signs in court</td>
<td>.721</td>
<td>.198</td>
<td>.449</td>
<td>3.634</td>
<td>.001</td>
</tr>
</tbody>
</table>

* Dependent Variable: was bail granted?

NOTE: dependent variable ‘was bail granted’ (1=no 2=yes)
Indep. variables: ‘princ. charge’ (1=not serious-property, yea, other cr 2=serious-violence, drugs, b&c)
‘child-like’ signs (1=no 2=yes-child-like characteristics)

Table 35 suggests that both the seriousness of the principal charge before the court and the presence of ‘child-like’ signs in the court hearing independently predict the likelihood of being granted bail in cases which were contested by the Crown attorney. If the charge was less serious, the courts were more likely to grant bail, and in the very few cases in which the youth exhibited signs that appeared to be ‘child-like’ the courts allowed the youth to be released before trial.

Section III - The overall detention decision

Legal Index:

Turning now to the overall detention decision which is presumably a combination of the Crown’s decision and the court’s decision on contested cases, it appears that the
presence of factors of 'legal seriousness' as defined in Appendix C were related significantly to the overall decision to grant bail.

Table 36

<table>
<thead>
<tr>
<th>Factors of legal seriousness</th>
<th>0 through 2 serious</th>
<th>3 serious</th>
<th>4 thru 5 serious</th>
<th>6 or more serious</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>no</td>
<td>1</td>
<td>5</td>
<td>12</td>
<td>17</td>
<td>35</td>
</tr>
<tr>
<td>yes</td>
<td>28</td>
<td>28</td>
<td>19</td>
<td>8</td>
<td>83</td>
</tr>
<tr>
<td>Total</td>
<td>29</td>
<td>33</td>
<td>31</td>
<td>25</td>
<td>118</td>
</tr>
</tbody>
</table>

As factors of legal seriousness increased, so too did the probability that the accused would be denied bail. As Table 36 shows, almost all (96.6%) of the youth with between 0 and 2 factors of legal seriousness were granted bail. As factors of legal seriousness increase, there is a dramatic decline in the probability of being granted bail—where youth who had 6 or more factors of legal seriousness were granted bail in only about one-third (32%) of the cases.

NOTE: Chi-square=31.711, df=3, p<.001
Principal Charge

The overall detention decision was related to the principal charge (Table 37).

Table 37

<table>
<thead>
<tr>
<th>Principal Charge</th>
<th>was bail granted?</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>no</td>
<td>yes</td>
<td>Total</td>
</tr>
<tr>
<td>Violence</td>
<td>19</td>
<td>22</td>
<td>41</td>
</tr>
<tr>
<td></td>
<td>46.3%</td>
<td>53.7%</td>
<td>100.0%</td>
</tr>
<tr>
<td>Drugs</td>
<td>4</td>
<td>12</td>
<td>16</td>
</tr>
<tr>
<td></td>
<td>25.0%</td>
<td>75.0%</td>
<td>100.0%</td>
</tr>
<tr>
<td>Break and Enter</td>
<td>6</td>
<td>13</td>
<td>19</td>
</tr>
<tr>
<td></td>
<td>31.6%</td>
<td>68.4%</td>
<td>100.0%</td>
</tr>
<tr>
<td>Property, Other</td>
<td>6</td>
<td>36</td>
<td>42</td>
</tr>
<tr>
<td></td>
<td>14.3%</td>
<td>85.7%</td>
<td>100.0%</td>
</tr>
<tr>
<td>Total</td>
<td>35</td>
<td>83</td>
<td>118</td>
</tr>
<tr>
<td></td>
<td>29.7%</td>
<td>70.3%</td>
<td>100.0%</td>
</tr>
</tbody>
</table>

NOTE: Chi-square=10.427, df=3, p=.015, 1 minimum expected value less than 5(4.75). Pooling ‘serious’ cases (violence, drugs, b&e) vs. ‘less serious’ cases (property, yo, other cc); Fisher’s exact test p=.007

Keeping in mind that the principal charge serves as only a broad index in understanding the ‘seriousness of the charge’, Table 37 shows that a higher proportion (46.3%) of those alleged to have committed offences that included violence were detained. It appears that the courts are also detaining almost one-third (31.6%) of those youths accused to have committed the offence of break and enter, whereas few of those charged with a drug offence (25%) or other property offence (14.3%) were denied bail.
Prior Record

Consistent with previous research (Gandy 1992; Carrington et al 1988; Bookin- Weiner 1984)\textsuperscript{22}, this study also found a significant relationship between prior record and the overall decision to detain.

Table 38 - Percentage of cases where bail was denied as a function of the relationship between prior record and current charges

<table>
<thead>
<tr>
<th>is there a prior record?</th>
<th>not mentioned</th>
<th>no</th>
<th>yes</th>
<th>sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.2%(31)</td>
<td>27.5%(40)</td>
<td>54.8%(42)</td>
<td>(p=.015^*) (no vs. y) (p&lt;.001) (overall including not mentioned's)**</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>are priors related?</th>
<th>unrelated</th>
<th>related</th>
<th>(p=.025^*)(related vs. unrelated) (p=.008) (overall)**</th>
</tr>
</thead>
<tbody>
<tr>
<td>25.0%(8)</td>
<td>33.3%(12)</td>
<td>77.3%(22)</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>are priors recent?</th>
<th>recent (within 1 year)</th>
<th>recent (within 1 year)</th>
<th>(p=.025^*)(related vs. unrelated) (p=.008) (overall)**</th>
</tr>
</thead>
<tbody>
<tr>
<td>33.3%(9)</td>
<td>37.5%(8)</td>
<td>68.0%(25)</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>do priors include violence?</th>
<th>does not include violence</th>
<th>does include violence</th>
<th>n.s.</th>
</tr>
</thead>
<tbody>
<tr>
<td>42.9%(14)</td>
<td>58.3%(12)</td>
<td>62.5%(16)</td>
<td></td>
</tr>
</tbody>
</table>

*Fisher’s exact test  
**2 expected values less than 5(1.48)  
***2 expected values less than 5(3.62)

Table 38 shows that youth with a prior record were more likely to be denied bail than those without. As well, having a prior record which included offences related to the current charges before the court was also related to the overall bail decision.

If the prior record was identified as being related to the current charge(s) before the court, in most cases (77.3\%), bail was denied. However, if this information was not stated in
court or if it came out in court that the prior record was not related to the current charge(s), there was a greater likelihood that bail would be granted.

Bail Abuse & Previous Pre-trial Detention:

As mentioned in Section I, in most (72%) of the cases observed, information about the youth's history of bail abuse was not mentioned. However, the Crown contested most (90%) of the cases where information on bail abuse was mentioned. As Table 39 shows, when the issue of bail abuse was raised, the youth was detained in 20 of the 29 cases (69%), irrespective if the information indicated that the youth did not have a history of bail abuse. This finding of course is solely a reflection of whether the Crown contested release. When the Crown did not contest release, information was typically 'not mentioned'.

Table 39 - Percentage of cases where bail was denied as a function of information on bail abuse or pre-trial detention

<table>
<thead>
<tr>
<th>has accused history of bail abuse?</th>
<th>yes</th>
<th>no</th>
<th>not mentioned in court</th>
<th>sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>90.0%(10)</td>
<td>57.9%(19)</td>
<td>16.4%(73)</td>
<td>-no vs. yes - n.s.</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>-overall including not mentioned's p&lt;.001*</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>has youth previously been in pre-trial detention?</th>
<th>yes</th>
<th>no</th>
<th>not mentioned in court</th>
<th>sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>73.3%(15)</td>
<td>71.4%(7)</td>
<td>19.8%(81)</td>
<td>-no vs. yes - n.s.</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>-overall including not mentioned's p&lt;.001**</td>
</tr>
</tbody>
</table>

* 1 expected value less than 5(3.14)
** 3 expected values less than 5(2.17)

The same situation is true for information about previous pre-trial detention. Out of the 22 cases where the Crown mentioned information on the youth's record of previous pre-trial detention, the large majority (73.3%) of these were denied bail.

---

22 Frazier & Bishop (1985) did not find prior record to be related to the detention decision
Reverse Onus

As reported in Section I, the relationship between the onus in the case and the Crown decision was approaching significance. For the Crown - it appeared that where there was clear reverse onus in a case, the Crown was equally likely to contest or consent to bail release. But, if the case was a Crown onus or when information about the onus was not mentioned in court, the Crown seemed more likely to consent to release the youth. For the overall decision to grant bail, there was no relationship between the onus and the decision to grant bail (Table 40).

Table 40

<table>
<thead>
<tr>
<th>Who has the onus?</th>
<th>was bail granted?</th>
<th></th>
<th></th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>no</td>
<td>yes</td>
<td>total</td>
<td></td>
</tr>
<tr>
<td>reverse</td>
<td>27</td>
<td>51</td>
<td>78</td>
<td></td>
</tr>
<tr>
<td>Crown onus or not</td>
<td>8</td>
<td>26</td>
<td>34</td>
<td></td>
</tr>
<tr>
<td>mentioned</td>
<td>31.3%</td>
<td>68.8%</td>
<td>100.0%</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>35</td>
<td>77</td>
<td>112</td>
<td></td>
</tr>
</tbody>
</table>

NOTE: Chi-square(corrected)=.88, df=1, not significant.

Co-accused(s)

When the issue of having a co-accused(s) was raised in the bail hearing, it did not seem to influence the decision to grant bail. As Table 41 shows, a youth with or without a co-accused was equally likely to have their bail granted or denied. As well, the presence or absence of a co-accused in the courtroom for the bail hearing was not related to the court's decision, nor was the information that a co-accused(s) was an adult.
Table 41 - Percentage of cases where bail was denied as a function of information about a co-accused(s) in the case

<table>
<thead>
<tr>
<th>does one+ charges include one+ co-accused(s)?</th>
<th>yes</th>
<th>no</th>
<th>not mentioned</th>
<th>sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>37.0%(54)</td>
<td>35.0%(40)</td>
<td>4.8%(21)</td>
<td>no vs. yes (n.s.) p=.018 (overall including not mentioned's)</td>
</tr>
<tr>
<td>are one or more co-accused(s) present?</td>
<td>27.3%(33)</td>
<td>55.6%(18)</td>
<td>n/a</td>
<td>n.s.</td>
</tr>
<tr>
<td>is one+ co-accused(s) an adult?</td>
<td>50.0%(8)</td>
<td>34.4%(32)</td>
<td>20.0%(10)</td>
<td>n.s.</td>
</tr>
</tbody>
</table>

'Child-like' vs. 'Adult-like' Index

The index created to look at 'child' and 'adult' like factors (see Appendix B), did not show a significant relationship with the final bail decision.

Table 42

<table>
<thead>
<tr>
<th>How many factors are associated with being a child or adult?</th>
<th>was bail granted?</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>no</td>
</tr>
<tr>
<td>adult or no</td>
<td>4</td>
</tr>
<tr>
<td>child-like</td>
<td>25.0%</td>
</tr>
<tr>
<td>1 child-like</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td>30.0%</td>
</tr>
<tr>
<td>2 child-like</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>20.5%</td>
</tr>
<tr>
<td>3 or more child-like</td>
<td>14</td>
</tr>
<tr>
<td></td>
<td>42.4%</td>
</tr>
<tr>
<td>Total</td>
<td>35</td>
</tr>
<tr>
<td></td>
<td>29.7%</td>
</tr>
</tbody>
</table>

NOTE: Chi-square=4.309, df=3, not significant.

As Section I showed there was also no relationship between the Crown decision to contest bail and the presence of 'child' or 'adult-like' factors. It was hypothesized that one of the reasons that these factors did not influence the Crown is because much of Crown decision-making occurs outside of the actual bail hearing. However, the justice of the peace or
judge does see the young person before the court in the bail hearing, and as Section II showed, there was a relationship between being ‘child-like’ in court and the detention decision—where all youth deemed to be ‘child-like’ were released in the cases contested by the Crown. However, the infrequency of the occurrence of these signs may account for their lack of an overall impact.

**Demographic Factors in the Bail Decision:**

**Age**

As with the Crown’s decision, there were no differences in the detention decision based upon which age group the accused youth was in. This finding is consistent with previous research (Carrington, Moyer and Kopelman 1988; Bookin-Weiner 1984; Frazier & Bishop 1985).²³

![Table 43](image)

**Table 43**

<table>
<thead>
<tr>
<th>Relationship between age group and the decision to grant bail</th>
</tr>
</thead>
<tbody>
<tr>
<td>was bail granted?</td>
</tr>
<tr>
<td>no</td>
</tr>
<tr>
<td>age group</td>
</tr>
<tr>
<td>12&amp;13</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>14&amp;15</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>16+</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Total</td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>

NOTE: Chi-square = .101 df=2, not significant.
Other demographic and ‘appearance related’ factors

Table 44 - Percentage of cases where bail was denied as a function of demographic and appearance-related factors

<table>
<thead>
<tr>
<th></th>
<th>male</th>
<th>female</th>
<th>sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>gender</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>male</td>
<td>female</td>
<td></td>
</tr>
<tr>
<td></td>
<td>32.0%(97)</td>
<td>19.0%(21)</td>
<td>n.s.</td>
</tr>
<tr>
<td>perceived ethnicity?</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>black</td>
<td>other</td>
<td></td>
</tr>
<tr>
<td></td>
<td>38.5%(39)</td>
<td>24.4%(78)</td>
<td>n.s.</td>
</tr>
<tr>
<td>are there ‘adult-like’ signs?</td>
<td>yes</td>
<td>no</td>
<td></td>
</tr>
<tr>
<td></td>
<td>40.0%(5)</td>
<td>29.2%(113)</td>
<td>n.s.</td>
</tr>
<tr>
<td>are there ‘child-like’ signs?</td>
<td>no</td>
<td>yes</td>
<td></td>
</tr>
<tr>
<td></td>
<td>31.8%(110)</td>
<td>0%(8)</td>
<td>n.s.</td>
</tr>
<tr>
<td>how is youth dressed for court?</td>
<td>not dressed up</td>
<td>dressed up</td>
<td></td>
</tr>
<tr>
<td></td>
<td>31.7%(101)</td>
<td>17.6%(17)</td>
<td>n.s.</td>
</tr>
<tr>
<td>Does youth have facial hair? (males only)</td>
<td>yes</td>
<td>no</td>
<td></td>
</tr>
<tr>
<td></td>
<td>27.3%(22)</td>
<td>31.7%(82)</td>
<td>n.s.</td>
</tr>
<tr>
<td>is youth interested in proceedings?</td>
<td>not interested</td>
<td>very/moderately interested</td>
<td></td>
</tr>
<tr>
<td></td>
<td>43.3%(30)</td>
<td>24.7%(85)</td>
<td>n.s.</td>
</tr>
</tbody>
</table>

* Fisher’s exact test

None of these demographic and appearance related factors actually related to the outcome of the case. Some of this is likely because there was so little variation on some of these factors.

All of the 8 youths who exhibited ‘child-like’ signs were granted bail but this was not statistically significant. Most (86%) of the youths in this sample were not dressed up for court. However, this did not significantly impact the detention decision.

---

23 Wordes, Bynum & Corley (1994) found a relationship between age and the detention decision
There was a small proportion (21% or 22/104) of youths in the sample that had facial hair. However, this characteristic did not significantly relate to the detention decision.

Discussion of Demographic and ‘Child-like’ vs. ‘Adult-like’ factors

Again, given that differences among youth appearing at bail hearings were so small, the influence of the ‘appearance’ of the youth was not statistically relevant in the overall detention decision. Thus, there was very little in the way of ‘constructing’ the youth in the bail hearing—most of the information presented was in relation to strictly legal factors.

Support Index

The index which examined factors associated with support (Appendix B), was not significantly related to the detention decision.

Table 45

<table>
<thead>
<tr>
<th>How many factors are associated with support for this youth?</th>
<th>was bail granted?</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>no</td>
<td>yes</td>
<td>Total</td>
<td></td>
</tr>
<tr>
<td>0 support</td>
<td>7</td>
<td>18</td>
<td>25</td>
<td></td>
</tr>
<tr>
<td>1-2 support</td>
<td>14</td>
<td>26</td>
<td>40</td>
<td></td>
</tr>
<tr>
<td>3 or more support</td>
<td>14</td>
<td>39</td>
<td>53</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>35</td>
<td>83</td>
<td>118</td>
<td></td>
</tr>
</tbody>
</table>

NOTE: Chi-square = .847, df=2, not significant.
Parental support

The presence of a parent in court was not related to the overall bail decision.

Table 46 - Percentage of cases where bail was denied as a function of parental presence in court & supervision

<table>
<thead>
<tr>
<th>parent(s) present?</th>
<th>yes</th>
<th>no</th>
<th>sig</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>30.1%(83)</td>
<td>31.3%(32)</td>
<td>n.s.</td>
</tr>
<tr>
<td>does court hear that parent(s) is involved in youth's life?</td>
<td>30.0%(80)</td>
<td>54.5%(11)</td>
<td>n.s.</td>
</tr>
<tr>
<td>does court hear that parent(s) able to supervise youth?</td>
<td>28.6%(56)</td>
<td>55.0%(2)</td>
<td>(p=.055^*) (n.s.)</td>
</tr>
</tbody>
</table>

*Fisher's exact test

Living Arrangements & School

However, Table 47 shows that living at home was a factor taken into account in deciding whether or not to detain the accused youth, a finding similar to that presented in Section I with respect to the Crown's decision.

Table 47

<table>
<thead>
<tr>
<th>Does this youth live with their parent(s)/guardian(s)?</th>
<th>was bail granted?</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>no</td>
</tr>
<tr>
<td>no</td>
<td>14</td>
</tr>
<tr>
<td></td>
<td>60.9%</td>
</tr>
<tr>
<td>yes</td>
<td>20</td>
</tr>
<tr>
<td></td>
<td>24.7%</td>
</tr>
<tr>
<td>not mentioned</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>7.7%</td>
</tr>
</tbody>
</table>

Total: 35 | 82 | 117

29.9% | 70.1% | 100.0%

NOTE: Fisher's exact test (2-tail) \(p=.002\) (no vs. yes).
Chi-square = 14.628, \(df=2, p=.001\) (overall including not mentioned's);
Expected value less than 5 (3.89).
A very large proportion (75.3%) of youths who lived with their parent(s) were granted bail release, while just over one-third (39.1%) of youths who did not live with their parents were released on bail.

**School attendance**

Being in school was also related to the detention decision. For those accused youths who attended school regularly, just over half (55.6%) were granted bail. However, not being in school meant that detention was almost certain, with only 18.8% of youth being released.

**Table 48**

<table>
<thead>
<tr>
<th>was bail granted?</th>
<th>no</th>
<th>yes</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>is the youth in school?</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>no</td>
<td>13</td>
<td>3</td>
<td>16</td>
</tr>
<tr>
<td>yes</td>
<td>16</td>
<td>20</td>
<td>36</td>
</tr>
<tr>
<td>not mentioned</td>
<td>6</td>
<td>59</td>
<td>65</td>
</tr>
<tr>
<td>Total</td>
<td>35</td>
<td>82</td>
<td>117</td>
</tr>
</tbody>
</table>

NOTE: Fisher's exact test (2-tail) \( p = .017 \) (no vs. yes). Chi-square=37.000, \( df = 2, \ p > .001 \) (overall including not mentioned's); 1 expected count less than 5(4.79).

Similar to the Crown decision then, the overall detention decision was related to factors which were connected to the ties in the community for the youth—living at home and attending school.
Conclusions Section III:

The overall bail decision, then, is one which encapsulates both the Crown decision and the court’s decision on contested cases. In order to see which of the variables significantly predicts this decision, Table 49 presents an ordinary least squares regression analysis.

Table 49

<table>
<thead>
<tr>
<th>Independent Variables</th>
<th>Unstandardized Coefficients</th>
<th>Standardized Coefficients</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>B</td>
<td>Std. Error</td>
</tr>
<tr>
<td>(Constant)</td>
<td>2.587</td>
<td>.286</td>
</tr>
<tr>
<td>principal charge</td>
<td>-.155</td>
<td>.058</td>
</tr>
<tr>
<td>prior record</td>
<td>-6.641E-02</td>
<td>.065</td>
</tr>
<tr>
<td>does youth live with parent(s)?</td>
<td>.134</td>
<td>.071</td>
</tr>
<tr>
<td>in school?</td>
<td>5.961E-02</td>
<td>.092</td>
</tr>
<tr>
<td>does Crown contest release?</td>
<td>-.632</td>
<td>.065</td>
</tr>
</tbody>
</table>

**NOTE:**
- Dependent variable ‘bail granted’ (1=no 2=yes)
- independent variables:
  - ‘principal charge’ (1=‘not-serious’-property, ya, other or 2=‘serious’-violence, drugs, b&e)
  - ‘prior record’ (1=no or not mentioned 2=yes)
  - ‘does youth live with parent(s)?’ (1=no 2=yes or not mentioned)
  - ‘is youth in school’ (1=no 2=yes or not mentioned)
  - ‘does Crown contest?’ (1=no 2=yes)

**cases where information was ‘not mentioned’ appeared to be associated with cases in which the Crown was not contesting release, and therefore where the information was not read aloud in court.

Looking at Table 49 shows that, not surprisingly, the Crown’s decision is strongly and independently related to the decision to grant bail. If the Crown did not contest, bail was likely to be granted. The seriousness of the principal charge before the court also had independent effects on the overall bail decision. If the charge was not as serious (minor property offences, ‘other Criminal Code’ & YOA offences), the courts were more likely to
release the youth. The variable describing if the youth lived with parent(s) was
approaching significance ($p=.061$).

Thus, to summarize the variables relating to bail decision-making in youth court, it
appears that the Crown’s decision (which was independently predicted by the youth
having a prior record and regularly attending school) was most influential on the outcome
of the case. If the Crown consented to release, all youths were released on bail. In those
cases where the Crown contested release ($n=48$), the variables that predicted the court’s
decision were the seriousness of the charge and the presence of ‘child-like’ factors (in a
small number of cases, $n=4$). The end result of these decisions was the overall bail
decision, which was independently predicted by the Crown’s decision (based on prior
record and school) and the seriousness of the principal charge before the court. In
addition, living at home appears to be positively associated with being granted bail.

Thus, key court decision makers (Crown attorney, Justice of the Peace or Judge)
are factoring in a number of different issues in determining bail eligibility, which relate to
the youth’s criminal past and their present social circumstances. These are; legal variables
(principal charge and prior record) which relate to the secondary ground for bail detention
(protection of society); and variables which may show ties to the community (being in
school and living with parents) which presumably relate to the primary ground for
detention (attending trial). Attending school and living with parents may also be
interpreted as ‘youthful’ characteristics, which are associated with a conventional
upbringing. In addition to these variables, rare cases of ‘child-like’ behaviour in court had
independent effects on the court’s decision for cases contested by the Crown.

82
Section IV: Conditions of Release

While being ‘youthful’ does not seem to benefit an accused, except in a small number of cases where the youth acts ‘child-like’ in court, it appears that age does come into play in the kinds of conditions that are placed upon youth who are released on bail. Many of the conditions placed on these individuals appear to be connected to their age.

In this sample, 81 (68.6%) youths were released on bail. Out of 81, all were given conditions of release. In examining whether there was a connection between the age of the youth and the kind of condition placed on the bail release order, these data show that ‘younger’ youths (those aged 12-15) were more likely to be given conditions such as a curfew order, or an order to reside with a particular person.

Table 50

<table>
<thead>
<tr>
<th></th>
<th>curfew</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>no</td>
<td>yes</td>
<td>Total</td>
</tr>
<tr>
<td>age group</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12-15</td>
<td>11</td>
<td>16</td>
<td>27</td>
</tr>
<tr>
<td></td>
<td>40.7%</td>
<td>59.3%</td>
<td>100.0%</td>
</tr>
<tr>
<td>16+</td>
<td>30</td>
<td>15</td>
<td>45</td>
</tr>
<tr>
<td></td>
<td>66.7%</td>
<td>33.3%</td>
<td>100.0%</td>
</tr>
<tr>
<td>Total</td>
<td>41</td>
<td>31</td>
<td>72</td>
</tr>
<tr>
<td></td>
<td>56.9%</td>
<td>43.1%</td>
<td>100.0%</td>
</tr>
</tbody>
</table>

Note: Fisher's exact test (2-tail) p=.049

When controlling for the youth's prior record, for those youth without a prior record, there were significant differences between groups, where higher proportions of the younger group were given curfew conditions (Table 51).
Table 51 - Percentage of cases given curfew as a bail condition as a function of prior record by age group

<table>
<thead>
<tr>
<th>prior record</th>
<th>age group</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>12-15</td>
<td>16+</td>
<td></td>
<td>sig.</td>
</tr>
<tr>
<td>prior record</td>
<td>66.7%(3)</td>
<td>38.5%(13)</td>
<td></td>
<td>n.s.</td>
</tr>
<tr>
<td>no prior record or not mentioned</td>
<td>58.3%(24)</td>
<td>29.6%(27)</td>
<td></td>
<td>p=.051* (n.s.)</td>
</tr>
</tbody>
</table>

*Fisher’s exact test

For the few youth who were granted bail and who did have a prior record, there were no significant differences between age groups in the likelihood of being given a curfew or not. This may be partially explained by the low numbers of youth who fell into this category.

The other condition of bail that seemed to be related to the youth’s age is the order to ‘reside’ with a particular person until the trial.

Table 52

<table>
<thead>
<tr>
<th>reside order given?</th>
<th>no</th>
<th>yes</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>age group</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12-15</td>
<td>4</td>
<td>23</td>
<td>27</td>
</tr>
<tr>
<td></td>
<td>14.8%</td>
<td>85.2%</td>
<td>100.0%</td>
</tr>
<tr>
<td>16+</td>
<td>22</td>
<td>23</td>
<td>45</td>
</tr>
<tr>
<td></td>
<td>48.9%</td>
<td>51.1%</td>
<td>100.0%</td>
</tr>
<tr>
<td>Total</td>
<td>26</td>
<td>46</td>
<td>72</td>
</tr>
<tr>
<td></td>
<td>36.1%</td>
<td>63.9%</td>
<td>100.0%</td>
</tr>
</tbody>
</table>

NOTE: Fisher’s exact test (2-tail) p=.005

As Table 52 reveals, the majority (85.2%) of the younger age group who were released were given reside orders, compared to just over half (51.1%) of the older group.
Controlling for prior record, Table 53 shows that even for those youth without a prior record, the younger group was still more likely to receive a reside order as part of their release conditions. However, for youth with a prior record, there were no significant differences between groups, but again, the number of youth with a prior record that fell into this category was quite small (n=21).

**Table 53 - Percentage of cases given a reside order as a bail condition as a function of prior record by age group**

<table>
<thead>
<tr>
<th></th>
<th>age group</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>12-15 years</td>
<td>16+</td>
<td>sig.</td>
<td></td>
</tr>
<tr>
<td>prior record</td>
<td>100%(3)</td>
<td>69.2%(13)</td>
<td>n.s.</td>
<td></td>
</tr>
<tr>
<td>no prior record or not mentioned</td>
<td>83.3%(24)</td>
<td>48.1%(27)</td>
<td>p=.018*</td>
<td></td>
</tr>
</tbody>
</table>

*Fisher's exact test

The potential for supervision, by way of curfews or reside orders, seems to be more of a concern for 'younger' youths. Therefore, it seems that rather than age being a construct of leniency or harshness by the courts, the court's concern themselves with age in relation to the level of supervision that 'younger' youths have available to them.
What kinds of conditions were generally placed on the youth in this sample?

Table 54- The most frequently placed conditions on the 81 youth granted bail

<table>
<thead>
<tr>
<th>Bail condition</th>
<th>number</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>reside with...surety, parent, etc.</td>
<td>51</td>
<td>62.9%</td>
</tr>
<tr>
<td>obey house rules</td>
<td>46</td>
<td>56.7%</td>
</tr>
<tr>
<td>no keeping company of...co-accused, victims, etc.</td>
<td>46</td>
<td>56.0%</td>
</tr>
<tr>
<td>curfew</td>
<td>34</td>
<td>41.9%</td>
</tr>
<tr>
<td>not in possession of....weapons.. non-medically</td>
<td>29</td>
<td>35.8%</td>
</tr>
<tr>
<td>prescribed drugs</td>
<td></td>
<td></td>
</tr>
<tr>
<td>boundary restrictions</td>
<td>26</td>
<td>32.0%</td>
</tr>
<tr>
<td>attend school</td>
<td>18</td>
<td>22.2%</td>
</tr>
<tr>
<td>house arrest</td>
<td>15</td>
<td>18.5%</td>
</tr>
<tr>
<td>obtain counselling</td>
<td>10</td>
<td>12.3%</td>
</tr>
</tbody>
</table>

Of these conditions, some were distinctly related to issues of the control and supervision of the youth. For example, ‘obeying the rules of the house’ was given in the majority (56.7%) of the cases. The courts are presumably attempting to empower parent(s) or parent-figure(s) to be able to control youths at home, thereby assuring that there is some form of supervision on them until the trial date. Similarly, reside orders fulfill the court’s concern that the youth is accounted for by a parent or someone in authority until the next court appearance. As well, the condition of attending school everyday may satisfy the court that the youth is under the supervision of school officials, however, one of the issues that may arise from this kind of condition is the potential that the influence of peers or other circumstances at school may have lead to the alleged charges in the first place.
Conclusions:

Bail hearings have been characterized as the "...undeveloped third world of criminal law" (Nasmith 1990). Each detention decision exposes deep conflicts between a deliberate system of justice, and the pressure to respond immediately to threats to the social order (Miller & Guggenheim 1990).

The detention of youths accused of offences even further exacerbates the seriousness of this stage in the criminal process. Statistics from 1993-94 show that on average, in the majority of the provinces of Canada, at least one in five young persons held in custody on any given day are in detention, not serving a court-imposed sentence (Moyer 1996). Clearly, many young people are unable to convince the courts that they should be released on bail. This is not surprising since the grounds for detention put young people at a severe disadvantage compared with adults, since they are dependent on others to prove to the court that they will attend trial and not commit further offences. As stated by Bookin-Weiner "...most juveniles must rely on their parents for food and shelter. They have neither the financial or legal ability to guarantee their own appearance...juveniles dependence on their families is a key element of their social and legal lives" (Bookin-Weiner 1984).

However, the Young Offenders Act guarantees that the special needs of youth will be upheld in the criminal process because of their age. Despite this recognition, this study has highlighted the fact that the special status afforded to young people under the Young Offenders Act does not play into the decisions made in youth court bail hearings. Indeed, provisions of the Young Offenders Act which have been written to protect young people in pre-trial detention hearings (such as the 'Responsible Person' provision S. 7.1) are not
even raised in court. Instead, legal variables account for much of the decision-making in youth court bail hearings (such as prior record and the 'seriousness' of the offence as examined through the principal charge). To a lesser extent other variables which assume 'ties to the community' for young people—attending school and living at home—are the other relevant pieces of information about a youth's history that relate to detention decisions. While these variables do speak to the ties to the community an accused youth has, they are also intertwined into a social construction of the 'youth' which assumes parental supervision at home and a structured and supervised life outside of the home and in school. Beyond these constructions of the youth in bail hearings, very little discussion occurs in these hearings about the principles of the Young Offenders Act and how the special needs or special status of youth may reflect upon the case at hand.

Moreover, the bail hearing itself is presumed to be the venue where information about the youth's case is to come out in front of the justice of the peace or judge so that an impartial decision can be made about the detention or release of the young person. However, consistent with previous research, this study suggests that this process is supplanted by the decision of the Crown attorney, who accounts for the large part of decision-making in bail hearings. As stated earlier, research suggests that the Crown recommendation is heavily influenced by the police recommendation in the show cause report (Wortley & Kellough 1998; Hucklesby 1998). This calls into question the accountability of decision-making which is in large part based upon the construction of the accused youth by other criminal justice agents, most notably the police, and the probability that decisions are being made executive rather than judicially (Hucklesby 1998).
What is also important to note is that there is very little evidence of the legislative provisions of the Young Offenders Act as influential in pre-trial detention hearings. Given that youth are generally dependent upon a responsible adult to be released, it is surprising to find that the ‘responsible person’ provision (s.7.1 of the YOA) was never raised in these hearings. There was also little evidence that the ‘youthfulness’ of these accused persons was being discussed at bail hearings. In the cases where the Crown had contested release and the youth showed distinctly ‘child-like’ factors, the judge or justice may have been affected by the accused’s youthfulness. But only in these rare examples did the ‘youth’ make a difference in court.

**How is the youth constructed and by whom?**

Clearly, the Crown constructs the case in a particular way when the case is being contested. The Crown only presents information about factors of legal seriousness (e.g. history of bail abuse) when a case is being made to contest release. There seems to be very little in court to counter the Crown’s recommendation. This study shows that defense counsel rarely presents a ‘case plan’ for the youth to the judge. In only 6 cases in this study was there a concrete case plan presented.

Overall, there are two separate occasions in court when a youth can be released on bail. The Crown may decide not to ‘show cause’ to the justice or judge, and subsequently all these youth are released. As this study has shown, the Crown opted not to show cause in almost 60% of the cases that came to court. In the remaining 40% of cases, the justice or judge may let the youth out on bail. Although, in only a minority of these remaining cases (just under 30%) did the justice or judge release the youth. Legal variables were the
main factors that were associated with these decisions, but factors relating to the
accused's 'youthfulness' appeared not to have any substantial impact on decisions made in
these bail hearings.
### Appendix A - bail coding sheet used in court observation

<table>
<thead>
<tr>
<th>Location &amp; Time</th>
<th>Address</th>
<th>Courtroom#</th>
<th>Date</th>
<th>AM</th>
<th>PM</th>
<th>Bail hearing</th>
<th>Bail review</th>
</tr>
</thead>
</table>

| Case Information | Docket/file# | Name | Perceptual Age | D.O.B. | Sex | Phase | Custody | Ethnicity | S/C
|-------------------|--------------|------|----------------|--------|-----|-------|---------|-----------|-------|

<table>
<thead>
<tr>
<th>(b) Docket/file#</th>
<th>Name</th>
<th>Age</th>
<th>Perceptual Age</th>
<th>D.O.B.</th>
<th>Sex</th>
<th>Phase</th>
<th>Custody</th>
<th>Ethnicity</th>
<th>S/C</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>(c) Lawyer:</th>
<th>Lawyer's Name</th>
<th>Is Present with Youth</th>
<th>Who Represents Youth @ Hearing</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Lawyer's Name</th>
<th>Is Present with Youth</th>
<th>Who Represents Youth @ Hearing</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>(d) Current Charge</th>
<th>What are the current allegations?</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Current Charge</th>
<th>What are the current allegations?</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>(e) Prior Record?</th>
<th># of Charges Read In</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Prior Record?</th>
<th># of Charges Read In</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>(f) Co-accused</th>
<th>Do one or more charges involve at least one other offender?</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Co-accused</th>
<th>Do one or more charges involve at least one other offender?</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Co-accused</th>
<th>Do one or more charges involve at least one other offender?</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Co-accused</th>
<th>Do one or more charges involve at least one other offender?</th>
</tr>
</thead>
</table>
what role does the v.o. being heard today allegedly have in the offence?
☐ equal to other offender(s)
☐ greater than other offender(s)
☐ less than other offender(s)
☐ not mentioned

(g) family:
are parent(s)/guardian(s) present in court?
☐ yes
☐ mother
☐ father
☐ guardian/foster
☐ no
☐ apparently
☐ not obvious/unclear
☐ parent(s) not waiting for case in courtroom
if no: why?
☐ parent(s) at work
☐ in hospital
☐ didn’t want to come
☐ no contact with y.o.
☐ deceased
☐ not mentioned
parent(s) involved in y.o.’s life?
☐ yes
☐ surety
☐ no
☐ not mentioned
☐ unlikely
do they wish to be involved in helping youth with offence?
☐ yes
☐ surety
☐ no
☐ not mentioned
☐ unlikely
are others present in court?
☐ yes
☐ no
☐ unclear
if yes, who?
☐ sibling(s)
☐ grandparent(s)
☐ other family
☐ other
☐ C.A.S.
☐ friend(s)

are parents/guardian(s) able to supervise v.o.?
(ie. work FT)
☐ yes
☐ no
☐ not mentioned
☐ unlikely
is parent/guardian(s) occupation mentioned?
☐ no
☐ yes:
is parent(s) on government assistance?
☐ Yes
☐ Family Benefits (Mother’s Allow)
☐ Disability (Workman’s Comp)
☐ Social Assistance (Welfare)
☐ U.I.C.
☐ no
☐ not mentioned

(h) living arrangements:
does youth normally live w/parent(s)?
☐ yes
☐ no
☐ no mention
if no: where does youth live?
☐ with other relative
☐ at friend’s house
☐ on the street
☐ ward of children’s aid/ch. welfare agency
(f) youth's appearance:
how is youth dressed for court?
- 'dressed up' - (suit, dress pants, dress shirt)
- 'neat' - (casual)
- dressed down (jeans, dress shirt...)
- other

are there 'non-child' or 'child' signs about the youth?
- no
- yes:
  - piercings______________________
  - goatee
  - moustache
  - peach fuzz
  - youth is a parent

other:_____________________________________

are there any signs about youth that show disinterest/disrespect in proceedings?
- no
- yes:

how interested is youth in proceedings? (can mark more than one)
- very interested
- moderately interested
- disinterested/looking around court
- defiant looking
- laughing (with friends)/aloof

(j) school:
- in school?
- no
- yes
  - name of school_____________________
  - last grade completed________________

- not mentioned
- most of the time
- sometimes
- once in a while
- is a student - but not presently registered
- unable to determine
- kicked out of school

how well is v.o. doing in school?
- excellent marks
- very good
- good in Special Ed.
- average
- passing all courses
- failing
- not mentioned

is youth involved in extra-curricular activities?
- yes
- no

-is youth involved in extracurricular activities?
- not mentioned
- likely not
- if yes: what?

is there other info presented from school?
- yes
- no

-if yes: what

(i.e. comments from teacher/principal/counselor)

Does v.o. have a learning disability?
- yes
- no

-is v.o. on or ever been on welfare?
- yes
- no

-is youth working?
- yes
- no

(k) bail hearing:
how long has youth been held?
Since: □day(s)□week(s)□month(s)

□ reverse onus?
□ crown onus?
□ unclear
□ not mentioned

if reverse onus, why?
□ committed indictable offence awaiting trial on another offence
□ breach of condition(s) of previous probation
□ offence under Narcotics Control Act
□ not a Canadian resident

does Crown contest the release of youth?
□ y yes: on what grounds?
□ primary (show up for court)
□ secondary (protect public)
□ n no: what are the conditions for release?
□ surety $________________
□ house arrest
□ obey rules of house
□ not in poss. of non-med pres.Narc
□ no keeping company of:
□ co-accused(s)
□ victim(s)

anyone known to have a criminal record
□ reside________________
□ attend school everyday____
□ notify of change of address____

final submissions
Crown:

Defense:

other comments (i.e. gendered information)
(J.D., C.K.S, or initials of witness etc. = info, presented/interpreted by ...)

witness #1
Defense: Examination:

surety? □ yes □ no □ unclear
release on youth's own recognizance?
□ yes □ no □ n/a
name________________________

relationship to accused?
□ mother
□ father
□ stepmother/stepfather
□ relative____________________
□ friend______________________
□ other_______________________
□ n/a

has known accused for how long? ______________________
□ all his/her life
□ more than 5 years
□ less than 5 years
□ n/a

what does witness do?
□ not mentioned
□ unemployed

on government assistance:
□ Family Benefits Allowance
□ Disability
□ Welfare
□ Unemployment Insurance

unskilled____________________
□ skilled____________________
□ managerial__________________
□ professional_________________
□ is employed (no mention of where)
who else lives with witness?
not mentioned
sibling(s) of this accused
other parent
other guardian
other ________________

plan of supervision
☐ no mention ☐ n/a

has witness ever signed bail before?
☐ y ☐ n ☐ no mention

circumstances: ____________________________
does youth have a history of bail escape?
☐ yes ☐ no
☐ not mentioned
☐ likely not

has v.o. been in pre-trial detention before?
☐ yes ☐ no
☐ not mentioned

does witness add conditions to bail?
☐ y ☐ n
what?

Judge/Justice - decision:
☐ bail granted
☐ bail denied

If granted: conditions:
surety and $ ____________
on house arrest
obey rules of house
carry bail papers @ all times
curfew: _________ to _________
not in possession of any non-med. pres. Narcotic
no keeping the company of:
c/o-accused(s)
victim(s)
anyone known to have crim. record

☐ reside
☐ attend school everyday
☐ notify change of address

If denied: on grounds:
☐ primary (show up for court)
☐ secondary (protect public)
☐ tertiary grounds
☐ not specified
comments: - judge:
Appendix B - inter-rater reliability

<table>
<thead>
<tr>
<th>level of discrepancy on observation between raters</th>
<th>#</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>total number of observations made together</td>
<td>569</td>
<td>100%</td>
</tr>
<tr>
<td>total agreement between raters</td>
<td>508</td>
<td>89.3%</td>
</tr>
<tr>
<td>slight differences between raters (e.g. not clear vs. yes, not mentioned vs. yes, perceived age within 2 years)</td>
<td>53</td>
<td>9.3%</td>
</tr>
<tr>
<td>total disagreement between raters (perceived age more than 2 years etc.)</td>
<td>7</td>
<td>1.2%</td>
</tr>
</tbody>
</table>
## Appendix C - Breakdown of Indices

<table>
<thead>
<tr>
<th>Index Name</th>
<th>Variables Included</th>
</tr>
</thead>
</table>
| **Legal**  | - youth placed in secure detention before trial (+)  
- most serious charge is either ‘violence’ or ‘drugs’ (+)  
- youth has a prior record (+)  
- 3 or more charges read into court record with respect to current allegations (+)  
- prior record has offences that are related to current offence (+)  
- prior record has offences that are recent (committed within the last year) (+)  
- prior offences include violence (+)  
- there is a co-accused in the current allegation(s) before court (+)  
- the co-accused is present in court on day of bail hearing - or also in custody (+)  
- the role of the youth in this offence is alleged to be greater than or equal to co-accuseds (+)  
- there is a reverse onus in the case (+)  
- the youth has a history of bail abuse (+)  
- the youth has previously been in pre-trial detention (+) |
| **Child vs. Adult** | - the real age of the youth is 12 through 15 years (+)  
- the perceived age of the youth is 15 years or less (+)  
- the youth lives with their parent (+)  
- the youth is in school (+)  
- the youth is involved in extra-curricular activities (+)  
- the youth exhibits other ‘child-like’ signs (i.e. crying on the stand) (+)  
- there is a co-accused who is in adult in the case (-)  
- the youth has facial hair (males only) (-)  
- the youth has been or is presently on welfare (-)  
- the youth exhibits one or adult-like signs (i.e. a parent) (-)  
- the youth is employed full-time(-) |
| **Support**  | - a parent/guardian is present at court, or the court hears why the parent/guardian could not come, and this is for legitimate reasons (i.e. at work) (+)  
- the parent/guardian is said to be involved in youth’s life (+)  
- the parent/guardian is able to supervise this youth (+)  
- there are others present in court for this youth (+)  
- the defence lawyer provides a case plan for this youth (+) |

+ means that this factor counted towards the overall index  
- denotes that this factor detracted from the overall index
Chapter Three – Sentencing hearings

Introduction

Section 3(c) of the YOA states that “young persons who commit offences require supervision, discipline and control but because of their state of dependency and level of development and maturity, they also have special needs and require guidance and assistance.” Based on this principle, one would expect that decisions made in youth court cases would relate to factors such as the offender’s age, maturity and development. There is evidence that, to some extent, this should be the case. As stated earlier in this thesis, research in developmental psychology has pointed out differences between ‘younger’ and ‘older’ adolescents in terms of their ability to participate in criminal justice proceedings. Younger youth are at a disadvantage in their understanding of the juvenile justice process (Scott and Grisso 1997). This is consistent with Canadian research on youths’ understanding of the legal process which has found younger youth at a significant disadvantage in terms of their ability to understand basic legal rights and in terms of general knowledge of the YOA (Abramovitch, Higgins-Biss and Biss 1993, Peterson-Badali and Koegl 1999).

Youth justice legislation has also provided legal demarcations relating to age. For instance, the former Juvenile Delinquents Act made distinctions based upon age in terms of incapacitation. Section 25 of the JDA stated that children under the age of 12 were not to be committed to an industrial school unless attempts were made to reform the youth in his home, a foster home, or a child welfare agency. Section 13(4) made it possible to put a youth over the age of 14 in pre-trial detention within the same facility as an adult at the discretion of key justice officials.
Even the current legislation, the YOA, divides young people up by age in certain circumstances. Based upon the most recent changes to the Young Offenders Act\(^1\), 16 and 17 year old youths who commit serious offences (murder, manslaughter, aggravated sexual assault), are to be presumptively transferred to adult court unless they can make a case to remain within the jurisdiction of youth court. Youths who are 14 and older, can also be transferred to adult court but the onus is on the Crown attorney to have them transferred to adult court. Finally, 12 and 13 year olds cannot, even for the most serious crimes, be transferred to adult court.

These legislative differences based upon age, coupled with the research on developmental differences among young people, suggests that there are important differences among youthful offenders in terms of age and responsibility. Young offenders are not a homogeneous group. The challenge is to understand how the differences among youth are accounted for in the youth justice process. As stated by Young (1989) "..[t]he special needs of young people are constantly alluded to, but the courts appear uncertain as to what exactly these special needs are and how it is that these special needs can be translated into a distinct penal philosophy" (Young 1989: 104). Therefore, the purpose of this chapter is to examine the role of age and 'youthfulness' as constructs in relation to the outcome of sentencing cases. As in Chapter 2 on bail hearings, this chapter will present findings on many of the same variables as they relate to the outcome of sentencing cases.

**Legal Context**

Dispositions under the *Young Offenders Act* attempt to strike a balance among the nature of the offence, the needs of the youth, and the interests of society (Beaulieu 1989).

\(^1\) Bill C-37 (1995); *Young Offenders Act* S.16
The Declaration of Principle (Section 3 of the *YOA*) sets out the guidelines for interpreting the *Act* and Section 20 (1) of the *YOA* provides a number of different sentencing options to judges when determining a sentence. These can range from an absolute discharge to a custodial sentence. However, when actually deciding upon youth court cases, the *Young Offenders Act* offers little guidance for judicial decision-making (Doob and Beaulieu 1992; Trépanier 1989; Young 1989). What the *Act* does provide decision makers, however, is a *framework* for formulating dispositions such as specifying the conditions that can be placed on probation orders, or setting out the circumstances which warrant the use of custody. Section 24 of the *Act*, which provides guidelines for the use of custody, is written with the *intent* of limiting the use of custody if possible. Thus, judges deciding cases under the *Young Offenders Act* are expected to use the conditions in Section 24, as well as the set of principles as outlined in Section 3 of the *YOA* when deciding on a sentence.

**Research Methods:**

Similar methods of data collection were used in determining sentencing outcomes as were outlined in Chapter 2 on bail hearings. From June to the end of August 1997, a research assistant and I observed sentencing hearings in various Toronto youth courts—the majority of observations being at Canada’s largest youth court in a downtown location of Toronto, Ontario. The data collection form\(^2\) used to gather information in court included demographic information (age, gender); legal factors (current offence(s), prior record); social factors (living arrangements, school involvement); personal characteristics (how the youth was dressed, how the youth acted in court); and information on representation in
court and the participation of offenders in their hearings (was there a joint submission? did the youth make a statement before sentencing?). Again, the focus of the research was the effect of age and apparent maturity on decision-making but in this case with respect to sentencing hearings. My expectation was that if a young person was described in more ‘youthful’ terms at the sentencing hearing (and within the pre-disposition report if used in a case), that he or she would be treated differently (more leniently) than would youths who were constructed in the hearing to be more ‘adult-like’ or mature. In all, 84 cases with a sentencing outcome were recorded. In addition, there were 17 cases in which a pre-disposition report was used in the observed sentencing hearing. Of these I was able to access 13 of the pre-disposition reports.

This chapter will be divided into four sections of analysis. The first will present findings on the factors relating to sentencing decisions based upon all of the variables recorded in the sample of observed youth court sentencing hearings. Section II will provide an analysis of references to the principles of sentencing in the YOA and the guidelines for the use of custody under S.24 of the YOA. Section III will examine the kinds of probation conditions placed on young offenders with a focus on the relationship between age and the nature of probation conditions. And Section IV will present an examination of predisposition reports with an analysis of how the information contained within them may relate to the youth’s age.

\(^2\) See Appendix A
The Sample of Observed Cases:

Just over one-quarter (27%) of the sample of observed cases received custody as the most significant disposition. This proportion of cases going to custody is consistent with Kowalski and Caputo's (1999) study which found that 25% of cases across Canada received a custodial sentence in 1995-96 and Moyer (1996) who found that 29% of cases received custody in 1993-94. In this sample, the majority of committals were open custody (17%) and the remainder were secure custody (11%) or a combination of both open and secure. Most (83%) of the custodial committals in this sample of observed cases were for 90 days or less which is consistent with custodial lengths reported elsewhere. For example, youth court statistics data for 1997-98 shows that 82.7% of cases heard in Ontario youth courts were for 90 days or less (Canadian Centre for Justice Statistics 1997).

Probation as a disposition was used alone or in combination with other sentencing options in the majority (74%) of cases. In over one quarter of these cases (27%) probation was the only disposition handed down.

More than half (53.5%) of the cases in this sample had as their principal charge, a minor property offence, offence under the Young Offenders Act, or 'other Criminal

---

3 Other cases had all the court information as well, but one had as its disposition 'time served' and thus was not counted as a sentence, one other was missing only the final sentence and was not counted.
4 The most significant disposition was coded as 'custody' (secure and open custody) over 'other' sanctions (community service orders, probation, fine, other intermediate sanctions, conditional discharges and absolute discharges).
5 The figures presented in this analysis (as well as the Kowalski and Caputo research (1999) and Moyer's research (1996)) differ from the figures presented in the youth court survey published by the Canadian Centre for Justice Statistics. The youth court survey defines a 'case' as charges that have the same date of first appearance in youth court. However, a set of charges presented on the same date in youth court does not always end up as the charges dealt with at disposition since some of the charges may have been withdrawn or stayed prior to disposition. The definition of a 'case' in this court observation research, Moyer (1996) and Kowalski and Caputo (1999) defines a case as those charges dealt with at the date of disposition. Therefore the total number of cases is smaller using this definition.
6 Such as escape custody, unlawfully at large, fail to comply with a probation order.
Just under one-third (30.2%) had as their principal charge an offence that included violence.

The majority of these cases (61%) dealt with youth who were 16 years of age or older, and most (86%) cases observed involved male offenders.

**Table 1 - Comparison of sample of youth sentencing cases observed between June and August 1997 to cases heard in Ontario youth courts (1997-1998) by principal charge, most significant disposition, age and sex**

<table>
<thead>
<tr>
<th>Principal Charge</th>
<th>Sentencing Sample</th>
<th>Ontario</th>
</tr>
</thead>
<tbody>
<tr>
<td>Violence</td>
<td>30.2%</td>
<td>24.5%</td>
</tr>
<tr>
<td>drug offences</td>
<td>9.3%</td>
<td>4.3%</td>
</tr>
<tr>
<td>break &amp; enter</td>
<td>7.0%</td>
<td>10.6%</td>
</tr>
<tr>
<td>other property</td>
<td>38.4%</td>
<td>34.7%</td>
</tr>
<tr>
<td>other Criminal Code</td>
<td>8.1%</td>
<td>16.8%</td>
</tr>
<tr>
<td>YOA offences</td>
<td>7.0%</td>
<td>9.0%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Most Significant Disposition</th>
<th>Sentencing Sample</th>
<th>Ontario</th>
</tr>
</thead>
<tbody>
<tr>
<td>Custody</td>
<td>27.3%</td>
<td>40.7%</td>
</tr>
<tr>
<td>Other</td>
<td>73.8%</td>
<td>59.3%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Age</th>
<th>Sentencing Sample</th>
<th>Ontario</th>
</tr>
</thead>
<tbody>
<tr>
<td>12-15 years old</td>
<td>38.8%</td>
<td>51.6%</td>
</tr>
<tr>
<td>16+ years old</td>
<td>60.2%</td>
<td>48.3%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Sex</th>
<th>Sentencing Sample</th>
<th>Ontario</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>86.0%</td>
<td>78.1%</td>
</tr>
<tr>
<td>Female</td>
<td>14.0%</td>
<td>21.9%</td>
</tr>
</tbody>
</table>

It appears from Table 1 that the sample of sentencing cases examined in this study are relatively comparable to data from Ontario from around the same time period. However, the sentencing sample involved more offences where the principal charge was violence or drugs. Custody was used in a smaller percentage of cases in the sentencing sample, which may be a result of the jurisdiction in which these data were collected. Given that the city of Toronto is a large metropolitan area, it is possible that there are more options

---

7 such as fail to comply with a disposition, fail to comply with an undertaking (this includes all cases except for YOA offences, drug offence and other federal offences).
8 the court in which the majority of cases were observed deals with all of the drug cases in the area, which may in part explain why there is a higher percentage of drug cases in the sentencing sample.
alternatives to custody available to judges than in other areas of the province. It is possible that, in having other options, judges in Toronto order custody in fewer youth court cases than do judges elsewhere.

Section 1 - Results:

In order to see if various factors grouped together made a difference in the sentencing decision, indices were created based upon 3 categories, ‘legal factors’, ‘child vs. adult’, and ‘support factors’ (see Appendix C for a breakdown of the variables included in each index). Each of the variables, if mentioned or observed in the sentencing hearing, counted as a ‘point’ toward the overall index. In the case of the ‘child vs. adult’ index ‘child-like’ factors counted positively towards the overall index, while any factors which constructed the youth in more ‘adult-like’ terms were subtracted from this index.

Legal Index:

First, Table 2 shows that there is a significant relationship between the index of legal factors and the resulting sentence.

---

9 For the analyses that follow, any cases with ‘missing information’ were deleted from the analysis.
The cases with 0 or 1 factors related to legal seriousness were less likely to be given a custodial sentence than were cases with 2 or more factors of seriousness. In fact, the courts were at least 3 times more likely to hand down a sentence of custody in cases where the young offender had 2 or more factors of legal seriousness over having only one or none at all.

**Breakdown of Legal Factors:**

**Pre-trial detention:**

An examination of each of the components of the legal index shows that certain individual legal variables were also related to the likelihood of receiving custody. For instance, being held in pre-trial detention had serious consequences for young people accused of offences in terms of their final sentence. As Table 3 shows, the large majority (87.8%) of youths who were not detained before trial did not end up receiving a custodial sentence, whereas just over half (51.4%) of those that were detained received a sentence.
other than custody. This finding is consistent with previous research on adults and youths (Fagan and Guggenheim 1996; McCarthy 1987; Koza and Doob 1975; Friedland 1965 and see Chapter 2).

### Table 3

**Relationship between the youth being held in pre-trial detention and the outcome of the case (custody or other)**

<table>
<thead>
<tr>
<th>Disposition</th>
<th>Custody</th>
<th>Other</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Held in pre-trial detention?</td>
<td>6</td>
<td>43</td>
<td>49</td>
</tr>
<tr>
<td></td>
<td>12.2%</td>
<td>87.8%</td>
<td>100.0%</td>
</tr>
<tr>
<td>Detained</td>
<td>17</td>
<td>18</td>
<td>35</td>
</tr>
<tr>
<td></td>
<td>48.6%</td>
<td>51.4%</td>
<td>100.0%</td>
</tr>
<tr>
<td>Total</td>
<td>23</td>
<td>61</td>
<td>84</td>
</tr>
<tr>
<td></td>
<td>27.4%</td>
<td>72.6%</td>
<td>100.0%</td>
</tr>
</tbody>
</table>

Fisher’s exact test *p* < .001

(NOTE: For all 2x2 tables in this analysis, Fisher’s exact test (2-sided) will be the reported statistic).

Table 4 below reveals that even when controlling for the possible influence of the principal charge, the relationship between pre-trial detention and custody still remains consistent in direction, but in most cases is not significant due in part to low numbers of cases.

### Table 4: Percentage of young offenders receiving custody as a function of being held in pre-trial detention by principal charge

<table>
<thead>
<tr>
<th>Principal Charge</th>
<th>Detained?</th>
<th></th>
<th>Sig</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Yes</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td>violence</td>
<td>50.0%(8)</td>
<td>17.6%(17)</td>
<td>n.s.</td>
</tr>
<tr>
<td>drugs</td>
<td>66.7%(3)</td>
<td>20.0%(5)</td>
<td>n.s.</td>
</tr>
<tr>
<td>break &amp; enter</td>
<td>50.0%(4)</td>
<td>0%(2)</td>
<td>n.s.</td>
</tr>
<tr>
<td>property, YOA or other Other CC</td>
<td>45.0%(20)</td>
<td>8.0%(25)</td>
<td>Fisher’s exact test <em>p</em> = .006</td>
</tr>
</tbody>
</table>
Principal charge:

Another important factor in sentencing a case to custody over an alternative sanction is likely to be the seriousness of the charge before the court. The Young Offenders Act is written in a way that implicitly highlights proportionality in sentencing - based upon the seriousness of the offence, the context of the offence, and the circumstances of the young person. As stated in section 24(1) of the Young Offenders Act;

...the youth court shall not commit a young person to custody unless the court considers a committal to custody to be necessary for the protection of society having regard to the seriousness of the offence and the circumstances in which it was committed and having regard to the needs and circumstances of the young person.

Research on the effects of the seriousness of the offence on the outcome of youth sentencing cases has shown mixed results. There is evidence that the seriousness of the offence is related to the outcome of the case but is not generally the primary factor that accounts for sentencing decisions. The effects of most serious charge occur in combination with other legal variables (prior record, number of charges at sentencing) and/or extra-legal variables (race, gender and age) (Kowalski and Caputo 1999; Staffensmeier, Ulmer and Kramer 1998; Lee 1995; Schissel 1994; Carrington and Moyer 1995; Doob and Meen 1993).

In this sample of cases, the principal charge\(^\text{10}\) before the court was not correlated with the youth receiving a sentence of custody. As Table 5 indicates, roughly equal proportions of young offenders received custody for different categories of offences.\(^\text{11}\)

\(^{10}\) The principal charge was computed by coding the first four charges in a case into 6 categories of offences; violence, drugs, break and enter, other property, other Criminal Code and YOA offences (in many observed cases though, there were fewer than four charges). Then, the principal charge was calculated by using the order of offences (consistent with Canadian Centre for Justice Statistics) where the 'most serious' charge trumped all other charges and was the unit counted.
Table 5

<table>
<thead>
<tr>
<th>Committed Crime</th>
<th>Custody</th>
<th>Other</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Violence</td>
<td>7</td>
<td>18</td>
<td>25</td>
</tr>
<tr>
<td></td>
<td>28.0%</td>
<td>72.0%</td>
<td>100.0%</td>
</tr>
<tr>
<td>Drugs</td>
<td>3</td>
<td>5</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>37.5%</td>
<td>62.5%</td>
<td>100.0%</td>
</tr>
<tr>
<td>B&amp;E</td>
<td>2</td>
<td>4</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>33.3%</td>
<td>66.7%</td>
<td>100.0%</td>
</tr>
<tr>
<td>Property &amp; CCC</td>
<td>11</td>
<td>34</td>
<td>45</td>
</tr>
<tr>
<td></td>
<td>24.4%</td>
<td>75.6%</td>
<td>100.0%</td>
</tr>
<tr>
<td>Total</td>
<td>23</td>
<td>61</td>
<td>84</td>
</tr>
<tr>
<td></td>
<td>27.4%</td>
<td>72.6%</td>
<td>100.0%</td>
</tr>
</tbody>
</table>

Chi-square=0.719, df=3, not significant.

It should be pointed out however that these offence categories may not be adequate proxies for understanding the 'seriousness' of a particular offence. Each offence category contains a broad range of offences that vary considerably in their 'seriousness'. For example, within the offence of 'assault' which is included in the 'violence' category are 3 levels of assault. The 'seriousness' could range then from a common assault which could be one person shoving another around --- to an aggravated assault. Thus, within each category of charge there could be a large degree of variation among actual offending behaviour, which may explain why there appears to be no effect of the principal charge on the use of custody in this sample of cases.

Prior Record:

There was, however, a significant relationship between a youth having a prior record and the likelihood of receiving a custodial disposition. The large majority (88.1%)  

---

Footnote:

11 Break and enter was separated out from other property offences since it is one of the more typical offences committed by youth and is generally seen as serious by the public and judiciary (see Moyer 1996, Gandy 1992).
of youths without prior records received sentences other than custody. However those with prior records had an almost equal chance of being placed in custody or not.

Table 6

<table>
<thead>
<tr>
<th>is there a prior record?</th>
<th>no</th>
<th>yes</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>5</td>
<td>37</td>
<td>42</td>
</tr>
<tr>
<td>Total</td>
<td>23</td>
<td>57</td>
<td>80</td>
</tr>
</tbody>
</table>

Fisher's exact test $p=.001$; The 4 cases where information on prior record was 'not mentioned' have been deleted from this analysis.

This relationship may be expected, since as was shown in Chapter 2, youths with prior records were more likely to be detained before trial. It is not surprising, therefore, to find that in this sentencing sample there was a relationship between being held pre-trial and having a prior record (see Table 7).

Table 7

<table>
<thead>
<tr>
<th>is there a prior record?</th>
<th>no</th>
<th>yes</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>32</td>
<td>13</td>
<td>45</td>
</tr>
<tr>
<td>Total</td>
<td>42</td>
<td>40</td>
<td>82</td>
</tr>
</tbody>
</table>

Fisher's exact test $p<.001$
As Table 7 shows, the large majority (73%) of those who were detained also had prior records in their cases.

There also appears to be a relationship between the outcome of a case and being held in pre-trial detention for those cases where a youth did have a record.\textsuperscript{12} As Table 8 reveals, looking only at youths who had prior records shows a greater likelihood of custody if they had been detained before trial (60%).

\textbf{Table 8}

\begin{tabular}{|c|c|c|c|}
\hline
\textbf{disposition} & & & \\
\hline
\textbf{pre-trial detention?} & \textbf{not detained} & \textbf{detained} & \\
\hline
\textbf{custody} & 3 & 15 & 18 \\
\textbf{other} & 10 & 10 & 20 \\
\hline
\textbf{Total} & 13 & 25 & 38 \\
\hline
\end{tabular}

Fisher's exact test $p = .043$

The relationship between prior record and custody was not significant for the most part when controlling for the type of the offence, but again, this was likely due to low numbers. As shown in Table 9 the relationship was still in the same direction; those with a prior record were more likely to be given a sentence of custody.

\textsuperscript{12} Note: for youth who did not have a record, the relationship was not significant due to low expected values but was in the same direction.
Table 9: Percentage of young offenders receiving custody as a function of prior record by principal charge

<table>
<thead>
<tr>
<th></th>
<th>Yes</th>
<th>No</th>
<th>sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>violence</td>
<td>50.0%(12)</td>
<td>8.3%(12)</td>
<td>Fisher's exact p = 0.069.</td>
</tr>
<tr>
<td>drugs</td>
<td>50.0%(4)</td>
<td>25.0%(4)</td>
<td>n.s.</td>
</tr>
<tr>
<td>break &amp; enter</td>
<td>40.0%(5)</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>property, YOA or OTHER CC</td>
<td>47.1%(17)</td>
<td>11.5%(26)</td>
<td>p = 0.014</td>
</tr>
</tbody>
</table>

* there were no young offenders whose principal charge was break & enter who did not have a prior record.

Despite the fact that prior record appeared to be related to the outcome of the case, having offences within the prior record which were identified in the sentencing hearing to be related to the current charges did not relate to the likelihood of being sent to custody. However, there appeared to be a significant relationship between a youth having a prior record which was recent or which included violence and receiving a custodial disposition.

Table 10: Percentage of young offenders receiving custody as a function of information on the nature of the prior record

<table>
<thead>
<tr>
<th>Relationship between the current charge and the nature of the prior record</th>
<th></th>
<th></th>
<th></th>
<th>sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>are priors related?</td>
<td></td>
<td></td>
<td></td>
<td>n.s. (overall)</td>
</tr>
<tr>
<td>related</td>
<td>45.5%(11)</td>
<td></td>
<td></td>
<td>related vs. unrelated</td>
</tr>
<tr>
<td>are priors recent?</td>
<td></td>
<td></td>
<td></td>
<td>n.s. (overall)</td>
</tr>
<tr>
<td>recent (within 1 year)</td>
<td>50.0%(12)</td>
<td></td>
<td></td>
<td>Fisher's exact test; p = 0.044 (recent vs. not recent)</td>
</tr>
<tr>
<td>are prior include violence?</td>
<td></td>
<td></td>
<td></td>
<td>p = 0.044 * (overall)</td>
</tr>
<tr>
<td>does include violence</td>
<td>80.0%(5)</td>
<td></td>
<td></td>
<td>Fisher's exact test; p = 0.032 (does not include vs. includes violence)</td>
</tr>
</tbody>
</table>

* The estimate of the statistical significance is likely to be exaggerated because there were 4 expected values less than 5 (2.24). If the minimum expected count is less than 5, this information will be noted from here on. The lowest expected value is indicated in parentheses.
There were few cases in which information on the relationship of the prior record to the current charges was stated. However, in the few cases where it came out that the prior record was recent or was one that included violence, there was a greater likelihood of the youth receiving a sentence of custody.

**Co-accused in case:**

One might expect that sentencing courts would deal with youth who commit crimes with others in a harsher manner for the purposes of general deterrence. The Supreme Court of Canada has suggested that a more punitive sentence may send a message to anyone else involved in the commission of an offence (R v. M.(J.J.)1993). In this sample of cases, half (50%) of the young offenders committed their crimes with one or more co-accused(s). Moyer's analysis of revised UCR data showed that 45% of all occurrences involving youth reported by police involved more than one suspect in 1992-93 (Moyer 1996: 53). In this sample of court observed cases, despite this information being mentioned in court, having a co-accused in the commission of the offence was not significantly related to the outcome of the case. In fact, equal proportions of those who committed offences with others and those who committed offences alone received custody (see Table 11). An analysis of the effects of having the co-accused present within the courtroom yielded similar results – there was no significant relationship between the court's knowledge of the co-accused being present in court, and the outcome of the case.

There was however, one piece of information about a co-accused that appeared to be related to judicial decision-making. Having a co-accused identified as an adult in the case was significantly related to the outcome of the case. While the number of times this
information was explicitly stated in court was low, it did seem to relate to the likelihood of a youth receiving a custody committal.

**Table 11: Percentage of young offenders receiving custody as a function of information about a co-accused in the case**

<table>
<thead>
<tr>
<th>does one+ charges include one+ other offenders?</th>
<th>percentage of cases receiving custody</th>
<th>Sig</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>28.6%(42)</td>
<td>25%(40)</td>
</tr>
<tr>
<td>No</td>
<td>22.2%(9)</td>
<td>23.1%(26)</td>
</tr>
<tr>
<td>are one+ co-accused(s) present?</td>
<td>62.5%(8)</td>
<td>0%(10)</td>
</tr>
<tr>
<td>is one+ co-accused(s) adult?</td>
<td></td>
<td>p=.007*</td>
</tr>
</tbody>
</table>

*Fisher's exact test (2-tailed)

NOTE: Out of the 42 cases where it was mentioned that there was a co-accused(s) in the case, in only 18 cases was it explicitly mentioned whether the co-accused was an adult or young offender.

As Table 11 suggests, if it was mentioned in court that a co-accused in the case was not an adult, all cases received a disposition other than custody. However, in cases where a co-accused(s) was identified as an adult, almost two-thirds (62.5%) of cases ended up in a custodial disposition. While the number of cases in which an adult was involved is low, it nevertheless appears to be consequential in the case. This may be a result of the perception that more serious kinds of offences are being committed with adult co-accused(s) or that the potential for further offending is more serious. Of the 8 cases which involved adults in this sample, 4 involved violence, 2 involved break and enters, and 2 involved minor property or administrative offences.

**Role of youth in offence**

When offences were committed with other young offenders, information about the role of the youth would sometimes be discussed at the sentencing hearing. One might presume that a greater level of involvement of the youth during the commission of an
offence will have some bearing on the severity of the disposition handed down. However, the results indicate that this information, when mentioned, was not related strongly or consistently to the likelihood of judges’ use of custody.

**Table 12**

<table>
<thead>
<tr>
<th>Disposition</th>
<th>Custody</th>
<th>Other</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Less than others</td>
<td>4</td>
<td>9</td>
<td>13</td>
</tr>
<tr>
<td>Equal to others</td>
<td>0</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Greater than others</td>
<td>2</td>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>Not mentioned</td>
<td>5</td>
<td>15</td>
<td>20</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>11</strong></td>
<td><strong>29</strong></td>
<td><strong>40</strong></td>
</tr>
</tbody>
</table>

*Not tested since expected values too small.*

**Age and demographic factors:**

**Child-like/Adult-like Index:**

The central focus of this research was an examination of the effects of more ‘youthful’ or more ‘adult-like’ constructions of the young offender in court in relation to the outcome of the case. While part of the philosophy of the *Young Offenders Act* is based upon recognizing differing levels of maturity and dependency for young offenders, there was no indication in this sample of cases that there were differences in the use of custody based upon an index of age related variables. This index included variables such as the youth’s chronological age, perceived age\(^{13}\), if they regularly attended school or

\(^{13}\) As assessed by the researchers in court
worked full-time, whether they lived with parent(s) or parent(s)-figures, among other factors (see Appendix C).

Table 13

<table>
<thead>
<tr>
<th>Number of Factors Relating to Being 'Child-like' or 'Adult-like'</th>
<th>Custody</th>
<th>Other</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adult or 0</td>
<td>8</td>
<td>11</td>
<td>19</td>
</tr>
<tr>
<td>Child-like</td>
<td>42.1%</td>
<td>57.9%</td>
<td>100.0%</td>
</tr>
<tr>
<td>1 Child-like</td>
<td>4</td>
<td>11</td>
<td>15</td>
</tr>
<tr>
<td>26.7%</td>
<td>73.3%</td>
<td>100.0%</td>
<td></td>
</tr>
<tr>
<td>2 Child-like</td>
<td>5</td>
<td>12</td>
<td>17</td>
</tr>
<tr>
<td>29.4%</td>
<td>70.6%</td>
<td>100.0%</td>
<td></td>
</tr>
<tr>
<td>3 or more Child-like</td>
<td>6</td>
<td>27</td>
<td>33</td>
</tr>
<tr>
<td>18.2%</td>
<td>81.8%</td>
<td>100.0%</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>23</td>
<td>61</td>
<td>84</td>
</tr>
<tr>
<td>27.4%</td>
<td>72.6%</td>
<td>100.0%</td>
<td></td>
</tr>
</tbody>
</table>

Chi-square=3.515, df=3, not significant.

Table 13 shows that there were no significant differences in the use of custody between groups who appeared to be more 'adult-like' or mature and those who appeared to have more 'child-like' factors.

The same holds true regarding the youth's chronological age group. Table 14 shows that there were no statistically significant differences in the use of custody based upon the young person's age group.
Table 14

<table>
<thead>
<tr>
<th>age group</th>
<th>custody</th>
<th>other</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>12&amp;13</td>
<td>.00</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>100.0%</td>
<td></td>
<td>100.0%</td>
</tr>
<tr>
<td>14&amp;15</td>
<td>7</td>
<td>19</td>
<td>26</td>
</tr>
<tr>
<td></td>
<td>26.9%</td>
<td>73.1%</td>
<td>100.0%</td>
</tr>
<tr>
<td>16+</td>
<td>16</td>
<td>35</td>
<td>51</td>
</tr>
<tr>
<td></td>
<td>31.4%</td>
<td>68.6%</td>
<td>100.0%</td>
</tr>
<tr>
<td>Total</td>
<td>23</td>
<td>60</td>
<td>83</td>
</tr>
<tr>
<td></td>
<td>27.7%</td>
<td>72.3%</td>
<td>100.0%</td>
</tr>
</tbody>
</table>

Chi-square=2.649, df=2, not significant.

While chronological age does not appear to be related to the use of custody in these cases, a plausible explanation for this may be that age is accounted for in earlier decisions in the criminal justice process. For example Lee (1995) found that probation officers’ decisions (in a U.S. jurisdiction) to refer cases into the formal court process or divert them informally were related both to legal variables (prior record, prior informal dispositions etc.) and the juvenile’s age controlling for other factors. Lee found that older youths (14-17) were more likely than younger accused persons to be referred into formal court processing. Similarly, Doob and Chan (1982) found that police decisions were related to legal variables (such as previous police contacts), the juvenile’s actions, and also the juvenile’s age; older youths (14-16) were more likely to be charged by police. Although, as the authors suggest, age is correlated with other factors such as the number of previous contacts with police (older offenders by virtue of being older are likely to have more previous contacts) and thus age may be subsumed within these other variables. For the
purposes of this study, it is possible then that age may come into the equation in decisions made before a case is brought to court.

When a case does come to court—at the sentencing hearing—other research has also found that age does not appear to relate to the outcome of the case. Moyer (1996), in her examination of Youth Court Survey data from 1993-94 found that use of custodial dispositions increases with age but that the differences are not as large as may be expected. However, as Moyer states, this analysis did not control for other factors such as prior record or the type of offence, which are themselves associated with demographic characteristics of the offender (Moyer 1996: 123-124). Kowalski and Caputo’s (1999) more recent analysis of youth court survey data for 1995-96 found that differences between age groups on the use of custody almost entirely disappeared when the number of prior convictions and the seriousness of the offence were controlled.

**Age groups and custody across Canada:**

The analysis presented thus far has presented data on age and the use of custody for the sample of observed cases in court (n=84). This sample provides an extensive amount of qualitative information on both chronological age and factors related to age as captured in the ‘child vs. adult’ index. But, in order to understand on a broader scale, how age might relate to the use of custodial dispositions, an analysis was undertaken of a sample of 43,936 youth court cases from 1996-97 compiled by Statistics Canada. This data set has information on the youth’s age, previous record history, offence and most serious disposition for the majority of the provinces in Canada.¹⁴ The following series of

¹⁴ It should be noted once again that these data do not correspond to the figures presented in the youth court surveys publications from the Canadian Centre for Justice Statistics. The youth court survey defines
tables (Table 15 through Table 20) provide an analysis of the relationship between the chronological age group of this sample of young offender cases heard in Canadian youth courts and the likelihood of receiving custody as a function of the presence or absence of a prior record and controlling for 12 categories of offences. Tables 15, 16 and 17 present these data for the three provinces in Canada with the largest youth populations; Ontario, Quebec and British Columbia. Tables 18, 19, 20 provide the same information for cases across Canada.

Table 15 shows the percentage of cases receiving custody as a function of age group and prior record.

---

a ‘case’ as a set of charges which enter into the court on the same date. These data look at a ‘case’ as a set of charges brought before a youth court on the same date of disposition. Thus, these data do not include any charges that have been withdrawn, stayed etc. They also, because they were focusing on the effect of criminal record, exclude YOA offences (which necessarily involved only those with records). Data for Nova Scotia are not included in this data set. Cases where the age was ‘unknown’ were deleted from this analysis.
Table 15 - Table showing percentage of cases receiving custody in 1996-97 as a function of age group by prior record for selected offences (involving violence) for selected major provinces (Ontario, Quebec and British Columbia)

<table>
<thead>
<tr>
<th>Offence</th>
<th>Age group</th>
<th>Ontario</th>
<th></th>
<th>Quebec</th>
<th></th>
<th>British Columbia</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Record</td>
<td>no record</td>
<td>Record</td>
<td>no record</td>
<td>Record</td>
<td>no record</td>
</tr>
<tr>
<td>assault w/weapon</td>
<td>13 &amp; under</td>
<td>47.1%(17)</td>
<td>14.3%(112)</td>
<td>66.7%(3)</td>
<td>13.3%(30)</td>
<td>33.3%(6)</td>
<td>0%(15)</td>
</tr>
<tr>
<td></td>
<td>14 &amp; 15</td>
<td>59.5%(84)</td>
<td>26.7%(210)</td>
<td>64.3%(14)</td>
<td>11.2%(107)</td>
<td>63.2%(19)</td>
<td>12.5%(48)</td>
</tr>
<tr>
<td></td>
<td>16+</td>
<td>69.9%(143)</td>
<td>30.0%(160)</td>
<td>36.9%(65)</td>
<td>15.2%(125)</td>
<td>67.6%(37)</td>
<td>23.9%(46)</td>
</tr>
<tr>
<td>sig.</td>
<td>p=0.081 (n.s.)</td>
<td>p=0.009</td>
<td>n.s.</td>
<td>n.s.</td>
<td>n.s.</td>
<td>p=0.062 (n.s.)</td>
<td></td>
</tr>
<tr>
<td>minor assault</td>
<td>13 &amp; under</td>
<td>56.1%(82)</td>
<td>10.6%(564)</td>
<td>0%(3)</td>
<td>5.8%(52)</td>
<td>37.5%(16)</td>
<td>9.8%(82)</td>
</tr>
<tr>
<td></td>
<td>14 &amp; 15</td>
<td>54.1%(364)</td>
<td>11.5%(889)</td>
<td>30.2%(53)</td>
<td>9.0%(155)</td>
<td>38.8%(85)</td>
<td>6.4%(204)</td>
</tr>
<tr>
<td></td>
<td>16+</td>
<td>47.0%(455)</td>
<td>11.5%(678)</td>
<td>38.4%(86)</td>
<td>7.6%(185)</td>
<td>37.4%(91)</td>
<td>4.7%(171)</td>
</tr>
<tr>
<td>sig.</td>
<td>p=0.077 (n.s.)</td>
<td>n.s.</td>
<td>n.s.</td>
<td>n.s.</td>
<td>n.s.</td>
<td>n.s.</td>
<td></td>
</tr>
<tr>
<td>robbery</td>
<td>13 &amp; under</td>
<td>53.3%(15)</td>
<td>30.4%(23)</td>
<td>66.7%(3)</td>
<td>21.1%(19)</td>
<td>50.0%(6)</td>
<td>18.8%(16)</td>
</tr>
<tr>
<td></td>
<td>14 &amp; 15</td>
<td>78.1%(64)</td>
<td>48.1%(79)</td>
<td>55.6%(27)</td>
<td>24.3%(115)</td>
<td>50.0%(28)</td>
<td>23.1%(52)</td>
</tr>
<tr>
<td></td>
<td>16+</td>
<td>89.2%(120)</td>
<td>61.9%(105)</td>
<td>76.6%(64)</td>
<td>40.2%(132)</td>
<td>77.3%(44)</td>
<td>57.8%(45)</td>
</tr>
<tr>
<td>sig.</td>
<td>p=0.001</td>
<td>p=0.012</td>
<td>n.s.</td>
<td>n.s.</td>
<td>p=0.016</td>
<td>p=0.043</td>
<td>p=0.001</td>
</tr>
<tr>
<td>other violence</td>
<td>13 &amp; under</td>
<td>60.0%(25)</td>
<td>19.3%(171)</td>
<td>50.0%(4)</td>
<td>18.4%(38)</td>
<td>0%(2)</td>
<td>9.5%(42)</td>
</tr>
<tr>
<td></td>
<td>14 &amp; 15</td>
<td>58.1%(105)</td>
<td>24.3%(267)</td>
<td>39.1%(23)</td>
<td>28.3%(92)</td>
<td>63.6%(22)</td>
<td>14.5%(55)</td>
</tr>
<tr>
<td></td>
<td>16+</td>
<td>57.8%(223)</td>
<td>24.0%(242)</td>
<td>54.9%(51)</td>
<td>28.8%(132)</td>
<td>51.0%(51)</td>
<td>15.1%(53)</td>
</tr>
<tr>
<td>sig.</td>
<td>n.s.</td>
<td>n.s.</td>
<td>n.s.</td>
<td>n.s.</td>
<td>n.s.</td>
<td>n.s.</td>
<td></td>
</tr>
</tbody>
</table>
Table 16 - Table showing percentage of cases receiving custody in 1996-97 as a function of age group by prior record for selected offences (property offences) for selected major provinces (Ontario, Quebec and British Columbia)

<table>
<thead>
<tr>
<th>Offence</th>
<th>age group</th>
<th>Ontario</th>
<th></th>
<th>Quebec</th>
<th></th>
<th>British Columbia</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Record</td>
<td>no record</td>
<td>Record</td>
<td>no record</td>
<td>record</td>
<td>no record</td>
</tr>
<tr>
<td>break &amp; enter</td>
<td>13 &amp; under</td>
<td>68.5%(54)</td>
<td>19.1%(225)</td>
<td>0%(7)</td>
<td>10.5%(57)</td>
<td>15.8%(19)</td>
<td>7.5%(53)</td>
</tr>
<tr>
<td></td>
<td>14 &amp; 15</td>
<td>73.2%(325)</td>
<td>21.6%(593)</td>
<td>42.9%(133)</td>
<td>15.6%(315)</td>
<td>46.3%(108)</td>
<td>10.7%(178)</td>
</tr>
<tr>
<td></td>
<td>16+</td>
<td>73.0%(549)</td>
<td>23.6%(602)</td>
<td>44.3%(388)</td>
<td>17.0%(471)</td>
<td>56.7%(141)</td>
<td>14.6%(158)</td>
</tr>
<tr>
<td></td>
<td>sig</td>
<td>n.s.</td>
<td></td>
<td>p=0.03</td>
<td>n.s.</td>
<td>p=0.002</td>
<td>n.s.</td>
</tr>
<tr>
<td>theft over</td>
<td>13 &amp; under</td>
<td>42.9%(7)</td>
<td>42.9%(14)</td>
<td>0%(2)</td>
<td>33.3%(3)</td>
<td>100.0%(3)</td>
<td>0%(4)</td>
</tr>
<tr>
<td></td>
<td>14 &amp; 15</td>
<td>66.7%(63)</td>
<td>22.5%(89)</td>
<td>39.1%(23)</td>
<td>6.7%(30)</td>
<td>55.6%(18)</td>
<td>9.1%(22)</td>
</tr>
<tr>
<td></td>
<td>16+</td>
<td>64.6%(82)</td>
<td>16.3%(98)</td>
<td>43.4%(53)</td>
<td>8.1%(74)</td>
<td>58.8%(34)</td>
<td>15.4%(26)</td>
</tr>
<tr>
<td></td>
<td>sig</td>
<td>n.s.</td>
<td></td>
<td>p=0.05</td>
<td>n.s.</td>
<td>n.s.</td>
<td>n.s.</td>
</tr>
<tr>
<td>theft under</td>
<td>13 &amp; under</td>
<td>32.6%(86)</td>
<td>5.5%(273)</td>
<td>37.5%(8)</td>
<td>3.8%(52)</td>
<td>16.0%(25)</td>
<td>0.9%(107)</td>
</tr>
<tr>
<td></td>
<td>14 &amp; 15</td>
<td>40.9%(364)</td>
<td>8.8%(741)</td>
<td>21.4%(98)</td>
<td>5.3%(207)</td>
<td>18.1%(149)</td>
<td>4.5%(269)</td>
</tr>
<tr>
<td></td>
<td>16+</td>
<td>38.2%(586)</td>
<td>6.9%(742)</td>
<td>24.7%(263)</td>
<td>8.9%(336)</td>
<td>24.5%(196)</td>
<td>3.9%(285)</td>
</tr>
<tr>
<td></td>
<td>sig</td>
<td>n.s.</td>
<td></td>
<td>n.s.</td>
<td>n.s.</td>
<td>n.s.</td>
<td>n.s.</td>
</tr>
<tr>
<td>possess stolen property</td>
<td>13 &amp; under</td>
<td>56.7%(30)</td>
<td>7.1%(84)</td>
<td>0%(2)</td>
<td>0%(4)</td>
<td>0%(5)</td>
<td>9.1%(22)</td>
</tr>
<tr>
<td></td>
<td>14 &amp; 15</td>
<td>58.4%(274)</td>
<td>15.2%(336)</td>
<td>40.0%(20)</td>
<td>8.6%(35)</td>
<td>44.8%(58)</td>
<td>7.5%(80)</td>
</tr>
<tr>
<td></td>
<td>16+</td>
<td>57.3%(459)</td>
<td>15.5%(431)</td>
<td>33.8%(74)</td>
<td>7.1%(70)</td>
<td>48.2%(114)</td>
<td>8.4%(107)</td>
</tr>
<tr>
<td></td>
<td>sig</td>
<td>n.s.</td>
<td></td>
<td>n.s.</td>
<td>n.s.</td>
<td>n.s.</td>
<td>n.s.</td>
</tr>
<tr>
<td>mischief/ damage</td>
<td>13 &amp; under</td>
<td>19.4%(36)</td>
<td>10.0%(140)</td>
<td>100.0%(1)</td>
<td>5.3%(19)</td>
<td>14.3%(7)</td>
<td>2.6%(38)</td>
</tr>
<tr>
<td></td>
<td>14 &amp; 15</td>
<td>43.4%(129)</td>
<td>5.9%(236)</td>
<td>15.8%(19)</td>
<td>1.7%(59)</td>
<td>18.5%(27)</td>
<td>2.6%(77)</td>
</tr>
<tr>
<td></td>
<td>16+</td>
<td>31.9%(135)</td>
<td>5.6%(178)</td>
<td>19.4%(62)</td>
<td>11.3%(80)</td>
<td>30.2%(53)</td>
<td>3.5%(85)</td>
</tr>
<tr>
<td></td>
<td>sig</td>
<td>p=0.01</td>
<td></td>
<td>n.s.</td>
<td>p=0.007</td>
<td>n.s.</td>
<td>n.s.</td>
</tr>
<tr>
<td>other property</td>
<td>13 &amp; under</td>
<td>45.8%(24)</td>
<td>7.8%(77)</td>
<td>0%(2)</td>
<td>5.6%(18)</td>
<td>25.0%(4)</td>
<td>6.3%(16)</td>
</tr>
<tr>
<td></td>
<td>14 &amp; 15</td>
<td>48.4%(124)</td>
<td>11.9%(226)</td>
<td>37.5%(16)</td>
<td>7.5%(40)</td>
<td>5.9%(17)</td>
<td>4.4%(45)</td>
</tr>
<tr>
<td></td>
<td>16+</td>
<td>40.7%(204)</td>
<td>10.4%(307)</td>
<td>25.8%(66)</td>
<td>9.6%(73)</td>
<td>32.1%(28)</td>
<td>6.1%(49)</td>
</tr>
<tr>
<td></td>
<td>sig</td>
<td>n.s.</td>
<td></td>
<td>n.s.</td>
<td>n.s.</td>
<td>n.s.</td>
<td>n.s.</td>
</tr>
</tbody>
</table>
Table 17 - Table showing percentage of cases receiving custody in 1996-97 as a function of age group by prior record for selected offences (other CC, drugs) for selected major provinces (Ontario, Quebec and British Columbia)

<table>
<thead>
<tr>
<th>offence</th>
<th>age group</th>
<th>Ontario record</th>
<th>Ontario no record</th>
<th>Quebec record</th>
<th>Quebec no record</th>
<th>British Columbia record</th>
<th>British Columbia no record</th>
</tr>
</thead>
<tbody>
<tr>
<td>other CC</td>
<td>13 &amp; under</td>
<td>46.7%(45)</td>
<td>12.9%(124)</td>
<td>25.0%(4)</td>
<td>0%(19)</td>
<td>37.5%(8)</td>
<td>20.0%(20)</td>
</tr>
<tr>
<td></td>
<td>14 &amp; 15</td>
<td>55.2%(277)</td>
<td>23.9%(444)</td>
<td>32.6%(43)</td>
<td>10.1%(99)</td>
<td>47.1%(85)</td>
<td>14.5%(76)</td>
</tr>
<tr>
<td></td>
<td>16+</td>
<td>49.1%(538)</td>
<td>19.8%(560)</td>
<td>22.3%(112)</td>
<td>4.9%(283)</td>
<td>37.8%(143)</td>
<td>7.9%(151)</td>
</tr>
<tr>
<td></td>
<td>sig.</td>
<td>n.s.</td>
<td>p=0.022</td>
<td>n.s.</td>
<td>p=0.093 (n.s.)</td>
<td>n.s.</td>
<td>n.s.</td>
</tr>
<tr>
<td>Drug related</td>
<td>13 &amp; under</td>
<td>33.3%(6)</td>
<td>15.2%(33)</td>
<td>0%(4)</td>
<td>10.3%(39)</td>
<td>50.0%(2)</td>
<td>11.1%(9)</td>
</tr>
<tr>
<td></td>
<td>14 &amp; 15</td>
<td>35.3%(68)</td>
<td>6.9%(202)</td>
<td>30.4%(56)</td>
<td>10.0%(331)</td>
<td>33.3%(18)</td>
<td>7.3%(55)</td>
</tr>
<tr>
<td></td>
<td>16+</td>
<td>32.9%(255)</td>
<td>10.2%(508)</td>
<td>30.1%(130)</td>
<td>8.5%(484)</td>
<td>19.4%(67)</td>
<td>7.9%(114)</td>
</tr>
<tr>
<td></td>
<td>sig.</td>
<td>n.s.</td>
<td>n.s.</td>
<td>n.s.</td>
<td>n.s.</td>
<td>n.s.</td>
<td>n.s.</td>
</tr>
</tbody>
</table>

Table 18 - Table showing percentage of cases receiving custody in 1996-97 as a function of age group by prior record for selected offences (involving violence) across Canada

<table>
<thead>
<tr>
<th>offence</th>
<th>age group</th>
<th>CANADA record</th>
<th>CANADA no record</th>
</tr>
</thead>
<tbody>
<tr>
<td>assault w/weapon</td>
<td>13 &amp; under</td>
<td>40.4%(52)</td>
<td>13.7%(226)</td>
</tr>
<tr>
<td></td>
<td>14 &amp; 15</td>
<td>59.6%(188)</td>
<td>18.9%(487)</td>
</tr>
<tr>
<td></td>
<td>16+</td>
<td>58.8%(388)</td>
<td>22.4%(473)</td>
</tr>
<tr>
<td></td>
<td>sig.</td>
<td>p=0.033</td>
<td>p=0.023</td>
</tr>
<tr>
<td>minor assault</td>
<td>13 &amp; under</td>
<td>39.0%(182)</td>
<td>8.7%(943)</td>
</tr>
<tr>
<td></td>
<td>14 &amp; 15</td>
<td>41.2%(779)</td>
<td>8.9%(1684)</td>
</tr>
<tr>
<td></td>
<td>16+</td>
<td>40.3%(977)</td>
<td>7.8%(1418)</td>
</tr>
<tr>
<td></td>
<td>sig.</td>
<td>n.s.</td>
<td>n.s.</td>
</tr>
<tr>
<td>robbery</td>
<td>13 &amp; under</td>
<td>47.1%(34)</td>
<td>21.4%(84)</td>
</tr>
<tr>
<td></td>
<td>14 &amp; 15</td>
<td>67.3%(199)</td>
<td>33.2%(310)</td>
</tr>
<tr>
<td></td>
<td>16+</td>
<td>81.8%(369)</td>
<td>50.3%(352)</td>
</tr>
<tr>
<td></td>
<td>sig.</td>
<td>p&lt;0.001</td>
<td>p&lt;0.001</td>
</tr>
<tr>
<td>other violence</td>
<td>13 &amp; under</td>
<td>42.9%(49)</td>
<td>16.9%(355)</td>
</tr>
<tr>
<td></td>
<td>14 &amp; 15</td>
<td>53.8%(260)</td>
<td>21.8%(611)</td>
</tr>
<tr>
<td></td>
<td>16+</td>
<td>55.3%(301)</td>
<td>23.1%(584)</td>
</tr>
<tr>
<td></td>
<td>sig.</td>
<td>n.s.</td>
<td>p=0.069 (n.s.)</td>
</tr>
</tbody>
</table>
Table 19 - Table showing percentage of cases receiving custody in 1996-97 as a function of age group by prior record for selected offences (property) across Canada

<table>
<thead>
<tr>
<th>offence</th>
<th>age group</th>
<th>CANADA</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>record</td>
</tr>
<tr>
<td>break &amp; enter</td>
<td>13 &amp; under</td>
<td>48.4%(186)</td>
</tr>
<tr>
<td></td>
<td>14 &amp; 15</td>
<td>58.0%(1025)</td>
</tr>
<tr>
<td></td>
<td>16+</td>
<td>58.8%(1811)</td>
</tr>
<tr>
<td></td>
<td>sig.</td>
<td>$p=0.024$</td>
</tr>
<tr>
<td>theft over</td>
<td>13 &amp; under</td>
<td>58.3%(24)</td>
</tr>
<tr>
<td></td>
<td>14 &amp; 15</td>
<td>60.0%(165)</td>
</tr>
<tr>
<td></td>
<td>16+</td>
<td>54.5%(266)</td>
</tr>
<tr>
<td></td>
<td>sig.</td>
<td>n.s.</td>
</tr>
<tr>
<td>theft under</td>
<td>13 &amp; under</td>
<td>23.9%(251)</td>
</tr>
<tr>
<td></td>
<td>14 &amp; 15</td>
<td>27.7%(132)</td>
</tr>
<tr>
<td></td>
<td>16+</td>
<td>28.0%(1772)</td>
</tr>
<tr>
<td></td>
<td>sig.</td>
<td>n.s.</td>
</tr>
<tr>
<td>possess stolen property</td>
<td>13 &amp; under</td>
<td>42.3%(71)</td>
</tr>
<tr>
<td></td>
<td>14 &amp; 15</td>
<td>48.6%(568)</td>
</tr>
<tr>
<td></td>
<td>16+</td>
<td>49.1%(1026)</td>
</tr>
<tr>
<td></td>
<td>sig.</td>
<td>n.s.</td>
</tr>
<tr>
<td>mischief/ damage</td>
<td>13 &amp; under</td>
<td>16.0%(81)</td>
</tr>
<tr>
<td></td>
<td>14 &amp; 15</td>
<td>30.5%(331)</td>
</tr>
<tr>
<td></td>
<td>16+</td>
<td>26.0%(557)</td>
</tr>
<tr>
<td></td>
<td>sig.</td>
<td>$p=0.026$</td>
</tr>
<tr>
<td>other property</td>
<td>13 &amp; under</td>
<td>33.3%(45)</td>
</tr>
<tr>
<td></td>
<td>14 &amp; 15</td>
<td>36.4%(253)</td>
</tr>
<tr>
<td></td>
<td>16+</td>
<td>32.3%(514)</td>
</tr>
<tr>
<td></td>
<td>sig.</td>
<td>n.s.</td>
</tr>
</tbody>
</table>
Table 20 - Table showing percentage of cases receiving custody in 1996-97 as a function of age group by prior record for selected offences (other CCC, drugs) across Canada

<table>
<thead>
<tr>
<th>offence</th>
<th>age group</th>
<th>CANADA</th>
<th>record</th>
<th>no record</th>
</tr>
</thead>
<tbody>
<tr>
<td>other CC</td>
<td>13 &amp; under</td>
<td>38.0%(108)</td>
<td>12.4%(226)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>14 &amp; 15</td>
<td>43.0%(572)</td>
<td>17.4%(867)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>16+</td>
<td>34.4%(1470)</td>
<td>10.9%(1644)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>sig.</td>
<td>p&lt;.001</td>
<td>p&lt;.001</td>
<td></td>
</tr>
<tr>
<td>drug related</td>
<td>13 &amp; under</td>
<td>33.3%(15)</td>
<td>10.7%(103)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>14 &amp; 15</td>
<td>31.7%(202)</td>
<td>7.6%(683)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>16+</td>
<td>26.6%(668)</td>
<td>8.4%(1324)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>sig.</td>
<td>n.s.</td>
<td>n.s.</td>
<td></td>
</tr>
</tbody>
</table>

It appears that only for the offence of 'robbery' is there a relatively consistent relationship between age and the likelihood of receiving custody across different jurisdictions\(^{15}\) regardless of record. As seen in Table 15 and Table 18, as age increased so too did the likelihood of a custodial sanction for the offence of robbery. This may be explained by the large range in the 'seriousness' of offences captured under the offence of 'robbery'. A robbery could feasibly include someone forcefully taking away another person's hat—all the way up to a robbery of a convenience store. Thus, judges may be responding to differences in the offender's age, in part, because of qualitative differences in the seriousness of the offence.

\(^{15}\) except Quebec where for robbery there were no significant differences in the likelihood of getting custody between age groups if there was a record.
For another relatively serious offence—assault with a weapon or causing bodily harm, older youths in Ontario without a record were dealt with more severely than were younger youth, 13 and under (Table 15). Looking at this same offence for all of Canada (except Nova Scotia) shows that there were also significant differences between age groups in the likelihood of receiving custody. Older youths were more likely to be given custodial sanctions whether they had a prior record or not (Table 18).

Finally there appear to be some differences across jurisdictions in the relationship of age to custody that are associated only with youths who have a prior record. In Ontario (Table 16) for example, youths with a prior record who were convicted of mischief or damage were most likely to receive custody if they were in the middle age group (14-15 years old). The same is true across Canada (Table 19), although this may be a function of Ontario comprising a large portion of all cases. In British Columbia, the likelihood of a custody sentence for the offence of break and enter (Table 16) was higher for older youths who had records. Across Canada, older youths (14 and up) were more likely to be given custody for a break and enter regardless of the presence of absence of a prior record (Table 19).

Thus it appears that only rarely, and only for certain offences (robbery being the most clear example) does the chronological age of a youth appear to be consistently related to the outcome of the case. However, for the most part, there are few substantial differences between age groups for most offences in these 3 provinces and across Canada in terms of the likelihood of receiving a sentence of custody. This finding is quite surprising in light of the developmental differences among youths who are governed by the
Based on these data, if a 13 year old and a 17 year old both commit a minor assault and neither has a prior record, they appear to be equally likely to be sentenced to custody.

**Breakdown of Demographic Factors in youth court sample:**

**Ethnicity:**

Because race/ethnicity statistics are not generally collected in youth court, we have very little data on the relationship of the offender’s ethnicity to the outcome of sentencing cases. In order to fill this gap in the research, both court observers assessed the most probable ethnicity of the young offender before the court. The categories were Caucasian, Black, South Asian, Aboriginal, Hispanic or other. Clearly this is not an appropriate measure of the *actual* ethnicity of the youth, but given that the focus of this research was to examine how information (both verbal and non-verbal) in court might affect decision-making, our perceptions of the ethnicity of the youth may have been consistent with the judges’ perceptions\(^{16}\) of the ethnicity of the youth (see Appendix B for inter-rater reliability). For the purposes of statistical analysis, race was coded into 3 major categories of ‘Black’, ‘Caucasian’ or ‘other’. Given that these were subjective assessments made by the researchers, these broad categories attempted to limit the capacity for error.
Table 21

<table>
<thead>
<tr>
<th>perceived ethnicity</th>
<th>custody</th>
<th>other</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>caucasian</td>
<td>8</td>
<td>35</td>
<td>43</td>
</tr>
<tr>
<td>black</td>
<td>10</td>
<td>10</td>
<td>20</td>
</tr>
<tr>
<td>other</td>
<td>5</td>
<td>15</td>
<td>20</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>23</strong></td>
<td><strong>60</strong></td>
<td><strong>83</strong></td>
</tr>
</tbody>
</table>

Chi-square = 6.814, df=2, p=.033
Caucasian/other (pooled) vs. Black; Fisher’s exact test p=.020

Table 21 suggests that young offenders who were perceived to be black by the researchers were more likely to be given custodial dispositions. This is consistent with other research on the effects of race (Staffensmeier, Ulmer and Kramer 1998; Schissel 1994 with respect to Native youths).

When controlling for prior record, the relationship between perceived ethnicity and the likelihood of receiving custody remains in the same direction but is not significant. It appears that a higher proportion of young offenders with prior records who were perceived to be black were given custodial sanctions. For young offenders without a record, the pattern appears to be in the same direction, with a slightly higher proportion of black youth being given custody (Table 22).

---

16 who were all apparently Caucasian
Table 22 - Percentage of youth receiving custody as a function of the perceived ethnicity controlling for prior record

<table>
<thead>
<tr>
<th>Perceived ethnicity</th>
<th>Black</th>
<th>Caucasian/Other</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>69.2% (13)</td>
<td>37.5% (24)</td>
<td>(n.s.) p = .091*</td>
</tr>
<tr>
<td>No</td>
<td>16.7% (6)</td>
<td>11.1% (36)</td>
<td>n.s.</td>
</tr>
</tbody>
</table>

*Fisher’s exact test

Also, when controlling for the nature of the principal charge, the relationship between perceived ethnicity and the likelihood of custody were not significantly related. However, for each offence type, the direction was the same. Table 23 shows that black youths who were charged with violent offences appeared to be more likely to be sentenced to custody than other youth in this sample of cases.

Table 23 - Percentage of youth receiving custody as a function of the perceived ethnicity controlling for principal charge

<table>
<thead>
<tr>
<th>Principal Charge</th>
<th>Black</th>
<th>Caucasian/Other</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Violence</td>
<td>80.0% (5)</td>
<td>15.8% (19)</td>
<td>p = .014*</td>
</tr>
<tr>
<td>Drugs</td>
<td>40.0% (5)</td>
<td>33.3% (3)</td>
<td>n.s.</td>
</tr>
<tr>
<td>Break &amp; Enter</td>
<td>50.0% (2)</td>
<td>25.0% (4)</td>
<td>n.s.</td>
</tr>
<tr>
<td>Property, YOA or OTHER CC</td>
<td>37.5% (8)</td>
<td>21.6% (37)</td>
<td>n.s.</td>
</tr>
</tbody>
</table>

*Fisher’s exact test

This may be the case though because there are differences among these groups in terms of such legal factors as being held in pre-trial detention. However, Table 24 reveals that when controlling for the effects of pre-trial detention there are consistent but statistically insignificant differences between groups in terms of their likelihood of receiving custody over other dispositions (Table 24).
Table 24 - Percentage of youth receiving custody as a function of perceived ethnicity holding pre-trial detention status constant

<table>
<thead>
<tr>
<th>pre-trial detention?</th>
<th>perceived ethnicity</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>black</td>
<td>caucasian/other</td>
</tr>
<tr>
<td>detained</td>
<td>66.7%(12)</td>
<td>39.1%(23)</td>
</tr>
<tr>
<td>not detained</td>
<td>25.0%(8)</td>
<td>10.0%(40)</td>
</tr>
</tbody>
</table>

Other demographic factors and appearance related factors:

Other variables which were examined in the court observation, some of which were demographic and others which related to how the youth 'appeared' in the court hearing, yielded no significant relationships when examined in relation to the likelihood of receiving a custodial disposition.
Table 25 - Percentage of cases receiving custody as a function of appearance and demographic variables in court

| Percentage of youth receiving custody as a function of demographic and appearance related factors in court |
|-------------------------------------------------|---------------------------------|----------------|
| gender                                         | male   | female                      | sig.       |
|                                                | 30.1%(73) | 9.1%(11)                        | n.s.       |
| adult-like signs?                              | YES   | NO                          |           |
|                                                | 25.0%(4) | 27.5%(80)                        | n.s.       |
| child-like signs?                              | NO    | YES                        |           |
|                                                | 26.6%(79) | 40.0%(5)                        | n.s.       |
| how was the youth dressed?                    | not dressed up | dressed-up                      | n.s.       |
|                                                | 26.7%(60) | 29.2%(24)                        |           |
| does youth have facial hair? (males only)     | YES   | NO                        |           |
|                                                | 44.4%(18) | 25.5%(55)                        | n.s.       |
| how interested is youth in sentencing proceedings? | not very interested | very / moderately interested | n.s.       |
|                                                | 29.4%(17) | 27.3%(66)                        |           |

There was a very small proportion (14%) of cases involving females in this sample.

As Table 25 shows, there were no statistically significant differences between males or females in relation to custody. This is consistent with Lee (1995), but other research (Kowalski and Caputo 1999; Staffensmeier, Ulmer and Kramer 1998; Schissel 1994) found gender effects on sentencing decisions. For instance, from their analysis of sentencing outcomes for adults (18 and over) between 1989-1992 in Pennsylvania, Staffensmeier et al (1998) found that females were sentenced in a less harsh manner than males. Canadian youth court observational research by Schissel (1993) analyzed sentencing outcomes in Alberta in 1986. Schissel found that for non-serious offences, females were dealt with more harshly than males, but for serious offences the opposite was true. According to youth court statistics in Canada for 1997-98, males were more likely
to be given sentences of custody. Thirty-six percent of cases involving males ended up in a custodial sentence compared to twenty-seven percent for females (Canadian Centre for Justice Statistics 1999: 91).

Beyond demographic variables such as gender, I also sought to measure how the youth appeared in court by recording information on variables such as whether the young offender exhibited ‘adult-like’ signs, ‘child-like’ signs, how the youth was dressed for court, whether the youth had facial hair, and whether or not the youth appeared interested in the sentencing hearing. As seen in Table 25 none of these variables yielded statistically significant findings when examined in relation to the likelihood of receiving a sentence of custody.

More importantly, variables where the youth did or had something that distinguished them by age —like exhibiting child-like characteristics (such as crying on the stand, or raising one’s hand to ask a question) or having adult-like attributes (such as being a parent, living in a common-law relationship), occurred quite infrequently. Other variables such as how the youth was dressed and whether or not they were interested in the proceedings had no apparent effect on whether the youth received custody. As with the bail sample (from Chapter 2) these young offenders at their sentencing hearings ‘appear’ to be a rather uniform group. Most (71%) dressed casually at their sentencing hearings —as average teenagers dress, and most (80%) appeared to be either very or moderately interested in the proceedings.
Support Index

I created one final index in order to examine if the level of support the court heard about for a youth might affect the outcome of the sentencing hearing. My hypothesis was that the more support a youth appeared to have from family or other supportive figures, the more likely the youth would be given a disposition other than custody. The variables that created this index included having a parent or parent-figure present at the sentencing hearing, the court hearing explicitly that the parent(s) was able to supervise the youth, the defense lawyer presenting a case plan for the youth, among others (see Appendix C).

Table 26

<table>
<thead>
<tr>
<th>Factors relating to support</th>
<th>Disposition</th>
<th></th>
<th></th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 support</td>
<td>Custody</td>
<td>9</td>
<td>16</td>
<td>25</td>
</tr>
<tr>
<td></td>
<td>Other</td>
<td>36.0%</td>
<td>64.0%</td>
<td>100.0%</td>
</tr>
<tr>
<td>1-2 support</td>
<td>Custody</td>
<td>8</td>
<td>26</td>
<td>34</td>
</tr>
<tr>
<td></td>
<td>Other</td>
<td>23.5%</td>
<td>76.5%</td>
<td>100.0%</td>
</tr>
<tr>
<td>3 or more support</td>
<td>Custody</td>
<td>6</td>
<td>19</td>
<td>25</td>
</tr>
<tr>
<td></td>
<td>Other</td>
<td>24.0%</td>
<td>76.0%</td>
<td>100.0%</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>23</td>
<td>61</td>
<td>84</td>
</tr>
<tr>
<td></td>
<td></td>
<td>27.4%</td>
<td>72.6%</td>
<td>100.0%</td>
</tr>
</tbody>
</table>

Chi-square=1.331, df=2, not significant.

Again, the data in this sample indicate that there were no significant differences among groups in terms of the levels of support a youth had and the likelihood of receiving custody over other dispositions.
Breakdown of support factors

Parents:

The presence of parents at sentencing hearings is said to be an important part of the process for the young offender. As stated in S.3(h) of the Young Offenders Act:

...parents have responsibility for the care and supervision of their children, and, for that reason, young persons should be removed from parental supervision either partly or entirely only when measures that provide for continuing parental supervision are inappropriate.

Schissel’s court observation research (1994) found that the presence of a parent in court (for Native youths) was critical in the outcome of a youth’s case in that not having a parent in court related to a more severe sentence for Native youths. When parents were present for Native youths there was a higher likelihood that the youth would plead not guilty.

In the sentencing cases observed, at least one parent or parent-figure was present in 58% of the cases. The presence of a parent in court, however, did not relate to the likelihood of receiving custody. As Table 27 shows, about a quarter of youth received a custodial disposition regardless of whether or not a parent was present.

In addition, while waiting for their sentencing hearing to be called, both young offenders and their supporters would generally sit in the body of the courtroom. An assessment was made by researchers as to whether or not the youth was sitting beside their parent or parent-figure. These data suggest that in about half of the cases the youth sat beside their parent or parent-figure and in the other half they did not.
Table 27 - Percentage of youth receiving custody as a function of parental presence in court, supervision, and living arrangements

<table>
<thead>
<tr>
<th>parent(s) present?</th>
<th>yes</th>
<th>no</th>
<th>sig</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>27.1%(48)</td>
<td>26.7%(30)</td>
<td>n.s.</td>
</tr>
<tr>
<td>is/are parent(s) sitting beside youth before hearing?</td>
<td>18.5%(27)</td>
<td>33.3%(24)</td>
<td>n.s.</td>
</tr>
<tr>
<td>does court hear that parent(s) is involved in youth's life?</td>
<td>26.7%(45)</td>
<td>20.0%(5)</td>
<td>n.s.</td>
</tr>
<tr>
<td>does court hear that parent(s) able to supervise youth?</td>
<td>24.0%(25)</td>
<td>22.2%(9)</td>
<td>n.s.</td>
</tr>
<tr>
<td>does youth live with parent(s)?</td>
<td>27.9%(54)</td>
<td>14.3%(14)</td>
<td>n.s.</td>
</tr>
</tbody>
</table>

The courts may or may not have been aware that the youth was sitting beside their parent or parent-figure. Regardless, as seen in Table 27, youths who were presumably sitting beside their parent(s)—indicating that there was at least some degree of a relationship between parent and youth—were no more likely receive a custodial disposition. The same holds true for the courts hearing that the parent(s) could supervise the youth, this information did not significantly relate to the outcome of the sentencing hearing. These findings are somewhat surprising in light of research which has examined what court personnel perceive to be important in considering sentencing youth. Sanborn (1996) found that the most frequently cited factor that ‘should’ be considered at sentencing was the ability of the family to supervise and assist with rehabilitation. The second most frequently cited factor was a delinquent record.

Living Arrangements:

Chapter 2 on bail hearings noted the significant relationship between living at home and being granted bail – youth who lived at home were more likely to be granted bail. In
these sentencing hearings however, there did not appear to be a statistically significant relationship between living at home and receiving a custodial sentence.

**Table 28: Percentage of youth receiving custody as a function of information about school, learning and employment status**

<table>
<thead>
<tr>
<th></th>
<th>yes</th>
<th>no/ not mentioned</th>
<th>sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>is the young offender in school?</strong></td>
<td>21.4%(56)</td>
<td>38.5%(26)</td>
<td>n.s.</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>n.s. (yes vs. no.)</td>
</tr>
<tr>
<td><strong>does the court hear that the youth has a learning disability?</strong></td>
<td>50.0%(6)</td>
<td>22.1%(68)</td>
<td>n.s.</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>n.s. (yes vs. no)</td>
</tr>
<tr>
<td><strong>is the young offender employed?</strong></td>
<td>26.3%(19)</td>
<td>24.6%(57)</td>
<td>n.s.</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>n.s. (yes vs. no)</td>
</tr>
</tbody>
</table>

**School:**

The relationship between being in school and bail being granted was also explored in Chapter 2, where a youth who attended school regularly was more likely to be granted bail. For sentencing hearings, the relationship between being in school and the likelihood of getting a custodial disposition was not a significant one, though the difference was in the same direction (lack of school attendance was associated with receiving a custodial sentence).

**Learning Disabilities:**

There is a large body of research which focuses on learning disabilities as one of the issues in need of attention in the youth justice system (Bell 1999, Scott and Grisso 1997, Jack and Ogloff 1997, Petersilia 1997, Winters 1997, Jarrelin et al 1994, Crealock 1991). The information that a youth had a learning disability or other disorder (such as Attention Deficit Disorder) was rarely mentioned in the sample of court cases, and in the few cases where it was raised, it did not relate to the outcome of the case.

**Employment:**
In only 28 cases was there explicit information in court about a youth's employment status. Again, there were no significant differences in the likelihood of getting custody for those who were said to be employed or not.

**Legal representation**

The final area of investigation of these sentencing hearings pertains to legal representation and the participation of youths and parents in court. As Table 29 shows, privately retained lawyers\(^\text{17}\) were used quite often at sentencing hearings. Of the privately retained lawyers' cases, almost half were for the categories of minor property, YOA offences, or 'other CCC' offences. Over one-third were for cases whose principal charge included violence. However, there were no significant relationships between the kind of legal representation in the case and the likelihood of receiving custody (Table 29).

**Table 29**

<table>
<thead>
<tr>
<th>dispostion</th>
<th>custody</th>
<th>other</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>who represents youth at hearing? duty counsel</td>
<td>6</td>
<td>16</td>
<td>22</td>
</tr>
<tr>
<td></td>
<td>27.3%</td>
<td>72.7%</td>
<td>100.0%</td>
</tr>
<tr>
<td>privately retained</td>
<td>17</td>
<td>43</td>
<td>60</td>
</tr>
<tr>
<td></td>
<td>28.3%</td>
<td>71.7%</td>
<td>100.0%</td>
</tr>
<tr>
<td>Total</td>
<td>23</td>
<td>59</td>
<td>82</td>
</tr>
<tr>
<td></td>
<td>28.0%</td>
<td>72.0%</td>
<td>100.0%</td>
</tr>
</tbody>
</table>

Chi-square(corrected) = .000, df=1, not significant.

\(^\text{17}\) which includes lawyers who were retained on a legal aid certificate.
Case plans:

Under the YOA young people are afforded a number of due process rights including the right to a lawyer. The right to a lawyer marked a significant shift in philosophy of the YOA over its predecessor the JDA. While due process rights were legislated to protect youth within the legal system, a great deal of research has called into question whether or not young people can fully exercise their rights. Most studies conclude that young people cannot meaningfully participate in the court process due to their lack of understanding of legal concepts (Peterson-Badali and Koepl 1998; Milner 1995; Abramovitch et al 1993; Peterson-Badali and Abramovitch 1992). This body of research highlights the importance of defense counsel’s role in defending young offender cases.

For example, studies which have examined the effectiveness of defense counsel in youth court have concluded that youth’s needs in court may not be adequately being served. First of all, there appears to be some disparity in the type of representation defense lawyers take on in youth court. Milne, Linden and Kueneman (1992) found that there was a large degree of variation in the roles that lawyers felt were appropriate in defending cases. The researchers found that most lawyers did not ‘fit’ into consistent roles of either advocate or guardian, but were a mix of these two extremes in defending youth cases. Second, research has pointed out that the type of legal representation a youth has may have consequences for the outcome of the case. Carrington and Moyer’s (1992) study of legal representation under the Juvenile Delinquents Act found that in one

\[18\.\text{In reference to the most appropriate role that should be taken by defense lawyers, the Ontario Law Society clearly supports the advocate role as most appropriate (Ontario Law Society }1981).\]
jurisdiction, the success rates of duty counsel were substantially lower than cases where private counsel was retained. The authors found that retained lawyers were more effective than duty counsel in having all charges terminated pre-plea as a result of negotiations with the prosecutor in the case. Finally, the role of defense counsel can be crucial in providing community alternatives to the court in order to avert custodial sentences. Macallair (1994) found that case 'advocates' in San Francisco (not necessarily defense lawyers) who provided case plans for juvenile cases were linked to a greater degree of deinstitutionalization of youth. By offering alternative information and case plans that were not available otherwise, a more individualized approach to cases resulted, and consequentially, fewer youth were institutionalized (Macallair 1994). Thus, information on alternative sanctions to custody is quite important for the offender's case. Judges also rate the importance of receiving information on community alternatives as very high in deciding cases (Hanscom 1988). Research suggests however that defense lawyers rarely present detailed dispositional plans for their young clients, nor adequate information on community alternatives to decision makers (Hanscom 1988).

In this study, there were only 10 cases out of 44 (23%) where a case plan was produced by defense counsel for the youth (Table 30)\textsuperscript{19}. In the remaining cases there was a joint submission by Crown and defense in which defense counsel did not present a case plan in the joint submission. Because of the low numbers of cases, it is difficult to assess the effect a case plan had on the likelihood of receiving custody. Nevertheless the

\textsuperscript{19} There was one case which had a case plan where the resulting sentence is missing.
direction indicates that there may be an association between case plans and dispositions other than custody.

Table 30

<table>
<thead>
<tr>
<th>Does defense provide a case plan?</th>
<th>Disposition</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Custody</td>
<td>Other</td>
<td>Total</td>
</tr>
<tr>
<td>no</td>
<td>12</td>
<td>63.6%</td>
<td>33</td>
</tr>
<tr>
<td></td>
<td>36.4%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>yes—case plan presented</td>
<td>2</td>
<td>7</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td>22.2%</td>
<td>77.8%</td>
<td>100.0%</td>
</tr>
<tr>
<td>joint submission and no case plan presented</td>
<td>8</td>
<td>33</td>
<td>41</td>
</tr>
<tr>
<td></td>
<td>19.5%</td>
<td>80.5%</td>
<td>100.0%</td>
</tr>
<tr>
<td>Total</td>
<td>22</td>
<td>61</td>
<td>83</td>
</tr>
<tr>
<td></td>
<td>26.5%</td>
<td>73.5%</td>
<td>100.0%</td>
</tr>
</tbody>
</table>

Chi-square = 2.760, df=2, not significant. ‘no vs. yes’ not significant.

NOTE: The number of cases with a joint submission does not correspond between Table 30 and 31. This is due to the fact that in cases where there was a joint submission for sentencing, defense counsel also put forward a case plan for the youth. Thus, of the 47 joint submission cases, most did not have a case plan from defense but a few of them had a case plan presented in the joint submission.

Joint Submission:

More than half (55%) of cases had a jointly submitted recommendation to the judge for sentencing and the judge agreed with the joint submission presented by Crown and defense in the majority (89%) of cases. This is consistent with Hanscom’s (1988) study which found that judges said they accepted joint submissions ‘all of the time’ or ‘usually’ (80.8% of cases).

Cases where there was joint submission tended to be less likely to result in custody (Table 31). This may simply be a function of the Crown agreeing to jointly recommend dispositions other than custody in certain non-serious cases.
Table 31

<table>
<thead>
<tr>
<th>is there a joint submission?</th>
<th>custody</th>
<th>other</th>
<th>Total</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>no</td>
<td>14</td>
<td>22</td>
<td>36</td>
<td></td>
</tr>
<tr>
<td></td>
<td>38.9%</td>
<td>61.1%</td>
<td>100.0%</td>
<td></td>
</tr>
<tr>
<td>yes</td>
<td>8</td>
<td>39</td>
<td>47</td>
<td></td>
</tr>
<tr>
<td></td>
<td>17.0%</td>
<td>83.0%</td>
<td>100.0%</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>22</td>
<td>61</td>
<td>83</td>
<td></td>
</tr>
<tr>
<td></td>
<td>26.5%</td>
<td>73.5%</td>
<td>100.0%</td>
<td></td>
</tr>
</tbody>
</table>

Fisher’s exact test (2-sided) $p = .043$

Participation of youth and parent(s) in sentencing hearing:

In just over 40% (34/81) of the cases, the judge asked the youth if s/he wanted to make a statement before sentencing. When asked to do so, the young offender made a statement more than half the time (19/34 cases). Making a statement or being asked to do so however, did not relate to the likelihood of receiving a sentence of custody.

Furthermore, parents made statements before sentencing in only 13 cases. This again however, had no relationship to the likelihood of receiving a sentence of custody.

Table 32 - Percentage of cases receiving a custodial disposition as a function of making a statement before sentencing in court

<table>
<thead>
<tr>
<th>does youth make a statement before sentencing?</th>
<th>yes</th>
<th>no/ or not asked</th>
<th>sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>42.1%(19)</td>
<td></td>
<td>23.8%(63)</td>
<td>n.s.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>does parent(s) make a statement before sentencing?</th>
<th>yes</th>
<th>no/ or not asked</th>
<th>sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>30.8%(13)</td>
<td></td>
<td>28.9%(38)</td>
<td>n.s.</td>
</tr>
</tbody>
</table>

There was a relationship between the youth making a statement before sentencing and the parent(s) doing so also. Of course this may be due to the dynamics of the courtroom,
where certain judges encouraged parties to speak - and thus both young offender and parent(s) made statements prior to sentencing (Table 33).

**Table 33**

<table>
<thead>
<tr>
<th></th>
<th>does parent make a statement before sentencing?</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>no</td>
</tr>
<tr>
<td>does youth make a</td>
<td></td>
</tr>
<tr>
<td>statement before</td>
<td>yes</td>
</tr>
<tr>
<td>sentencing?</td>
<td>33</td>
</tr>
</tbody>
</table>

Fisher’s exact test (2-sided) $p = .006$

**Conclusions of Section I**

From these data then, it appears that legal variables (pre-trial detention, prior record, co-accused who is an adult) and ethnicity are all related to the likelihood of youths in this sample receiving custody. In order to see if any of these variables significantly predicts the likelihood of custody, Table 34 presents a regression analysis.\(^20\)

---

\(^20\) a logistic regression analysis might have been seen as being a more appropriate test for the predictability of these variables since the dependent variable ‘custody’ is binary. However, because the sample size here is too small for logistic regression (under 100 cases), an ordinary least squares regression analysis was performed in order to best estimate the effects of the independent variables on custody.
Table 34

Ordinary Least Squares Regression analysis representing estimated effects of independent variables on the likelihood of receiving a custodial sentence a

<table>
<thead>
<tr>
<th>Independent variables</th>
<th>unstandardized Coefficients</th>
<th>standardized coefficients</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>B</td>
<td>Std. Error</td>
</tr>
<tr>
<td>(Constant)</td>
<td>2.831</td>
<td>.213</td>
</tr>
<tr>
<td>held in pre-trial</td>
<td>-.206</td>
<td>.099</td>
</tr>
<tr>
<td>detention</td>
<td></td>
<td></td>
</tr>
<tr>
<td>prior record</td>
<td>-.202</td>
<td>.099</td>
</tr>
<tr>
<td>adult co-accused</td>
<td>-.233</td>
<td>.147</td>
</tr>
<tr>
<td>perceived ethnicity</td>
<td>-.201</td>
<td>.105</td>
</tr>
</tbody>
</table>

a Dependent Variable: custody

NOTE: Dependent variable: 'custody' coded as (1=custody 2=other)

Independent variables: 'held in pre-trial detention' (1=not detained 2=detained)

'prior record' (1=no/ not mentioned 2=yes prior record)

'adult co-accused' (1=no/not mentioned/not applicable 2=yes-adult co-ac)

'perceived ethnicity' (1=caucasian/other 2=black)

Based upon the regression analysis, it appears that legal variables account for much of the predictability in the use of custody. Pre-trial detention and prior record both independently predict the likelihood of a sentence of custody, while the perceived ethnicity of the youth was approaching significance (p=.060).

Section II

Principles and goals of sentencing:

Another way in which 'youthfulness' as a factor in sentencing might be brought into sentencing hearings is through references to the principles and goals of sentencing relating to young offenders. An analysis of the cases in this sample shows that there were few cases (19/84 or 23%) in which reference was made to the principles of the Young Offenders Act or the general goals of sentencing. As seen in Table 35, an assessment was made of each statement made in court relating to goals and principles of sentencing. I
coded whether or not the principle or goal appeared to be stressed by the criminal justice agent as a 'relevant' factor in the case, 'not a relevant factor' in the case, or whether this could not be determined - 'neutral'. In Table 35 below, for each player in court, the first column shows how many times they made reference to the principle or goal, the next column shows how many of these references were 'neutral', followed by a column which shows how many of these were 'relevant' references, and the last column shows if the reference made appeared to limit the importance of the goal or principle ('not relevant'). The total number of references for each principle or goal is calculated in the second to last column to the far right, and the total number of cases where goals or principles were mentioned is calculated in the far right column (see Appendix D for more detailed breakdown of actual statements).
Table 35 - Frequency of references made to principles and goals of sentencing by key players in the courtroom in the 19 cases where reference was made

<table>
<thead>
<tr>
<th>Principal / goal of sentencing</th>
<th>Defense</th>
<th></th>
<th></th>
<th></th>
<th>Crown</th>
<th></th>
<th></th>
<th></th>
<th>Judge</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th>total # of refs</th>
<th>total # of cases</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td># refs</td>
<td>Ref</td>
<td>Not</td>
<td>rel</td>
<td># refs</td>
<td>Ref</td>
<td>Not</td>
<td>rel</td>
<td># refs</td>
<td>Ref</td>
<td>Not</td>
<td>rel</td>
<td>total</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>general deterrence</td>
<td>3</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>5</td>
<td>2</td>
<td>3</td>
<td>0</td>
<td>5</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>13</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>specific deterrence</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>5</td>
<td>4</td>
<td>1</td>
<td>0</td>
<td>9</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td>rehabilitation</td>
<td>3</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>3</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>9</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>public interest/public policy</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>6</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>public protection</td>
<td>0</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>4</td>
<td>3</td>
<td>1</td>
<td>0</td>
<td>5</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>general YOA principles</td>
<td>3</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>4</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>general sentencing principles</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>4</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>denunciation</td>
<td>0</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>2</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>young person's need to be held responsible for their behaviour</td>
<td>0</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>least restrictive measures (s. 24)</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>crime prevention</td>
<td>0</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td><strong>totals:</strong></td>
<td>14</td>
<td>6</td>
<td>7</td>
<td>1</td>
<td>17</td>
<td>10</td>
<td>7</td>
<td>0</td>
<td>25</td>
<td>23</td>
<td>2</td>
<td>0</td>
<td>56</td>
<td>out of 84 cases</td>
<td></td>
</tr>
</tbody>
</table>

What is most notable from the above table is that principles that are specific to young offenders (in S.3 of the YOA) were rarely raised in these sentencing hearings. In only 7 cases out of the total of 84 cases was there reference made to the principles of the YOA; 4 of these references pertained to general YOA principles and the remaining 3
references were about the use of least restrictive measures, crime prevention and the need for young persons to be held responsible for their behaviour. Thus, the specific principles that are to apply to young people appeared not to be raised at these sentencing hearings. However, issues regarding the public interest or public protection were raised more often. Out of the 84 cases, 10 cases had reference to public protection or public interest. There were a total of 11 references to these categories and the majority of these were brought up by the judge, followed by the Crown.

What is also quite important to note is that among the references to goals of sentencing, the majority of references that were made had to do with deterrence. Almost 40% (22/56) of references were made regarding either general or specific deterrence. Of these, the judge made the most reference to deterrence followed by the Crown. In addition, deterrence (specific or general) was raised in 12 of the 84 (14.3%) cases, which reveals the importance of deterrence as a factor in these sentencing hearings.

Cases that received custody:

As with the principles and goals of sentencing, another way that the court may be reminded that the offender before the court is a youthful offender is through references to Section 24 of the YOA which sets out the guidelines for the use of custody in youth courts. As stated earlier in this chapter, just over one quarter (27%) of these cases ended up in custody. The following legislative guidelines are relevant for decision makers when deciding whether to use custody (S.24 YOA):

---

21 In a study by Ouimet and Coyle (1991) which provided hypothetical sentencing case scenarios to prosecutors, judges, defense counsel and probation officers, prosecutors' sentencing preferences were found to be the most punitive of the 4 professional groups. In light of this study, these findings may in part be explained by the prosecutor's concern with public protection.
• custody should not be used unless it is necessary for the protection of the public having regard to the seriousness of the offence and the circumstances in which it was committed and having regard to the needs of the young person

• custody should not be used as a substitute for child welfare measures

• a young person who commits an offence that does not involve serious personal injury should be held accountable to the victim and society by way of non custodial dispositions when appropriate

• custody should only be imposed when all other alternatives to custody that are reasonable have been considered

• the youth court should consider a pre-disposition report unless it is deemed to be unnecessary

• where the youth court makes a custodial disposition they must state the reasons why other dispositions would not have been adequate

It would be fair to assume then, that when custody was being considered in a case, that some mention of the guidelines in Section 24 would come out in court. An analysis of the 25 cases\(^2\) where custody was almost certainly being considered suggests otherwise (see Appendix E for more detailed breakdown).

<p>| Table 36 - References to S. 24 of the Young Offenders Act in cases where custody was being considered |
|----------------------------------|---|---|---|---|</p>
<table>
<thead>
<tr>
<th>reference to S. 24 in observed court hearing?</th>
<th>No</th>
<th>Yes</th>
<th>somewhat</th>
<th>total:</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>19</td>
<td>2</td>
<td>4</td>
<td>25</td>
</tr>
<tr>
<td>is there reference to S.24 in the PDR?</td>
<td>13</td>
<td>0</td>
<td>-</td>
<td>13*</td>
</tr>
</tbody>
</table>

* although in 25 cases custody was being considered, I had access to only 13 PDRs.

Of the total 25 cases where custody was being considered, in only six cases was there some degree of reference to the guidelines in Section 24 of the YOA. S.24 was very explicitly mentioned in only two cases, in one case, defense counsel stated "...S.24 of the

\(^2\) 23/84 cases received custody and 2 other cases had a pre-disposition report but did not end up with a custodial sentence (n=25).
YOA says jail should be used as a last resort...[and the court should use] the least restrictive measures.” In the other case, defense counsel argued that “...any sentencing principles can be met by the maximum hours of community service.” Of these 25 cases, 2 were not given a custodial sentence. It appears then from these cases that the guidelines of S.24 are rarely being raised in youth courts when custodial sanctions are being considered.

Section III - Probation:

The analysis thus far has focused on the use of custody as the outcome measure in relation to age and the other variables recorded in the court. However, the absence of any relationships between age and custody may mean that there are other ways in which judges are responding to age-related issues when deciding on sentences for youth. Thus, I also analyzed the use of probation in relation to the youth’s age.

Section 23 of the Young Offenders Act states that the following conditions shall be included in all probation orders:

- keep the peace and be of good behaviour
- appear before the youth court as required

Judges have the option of placing other conditions on youth probation orders:

- that the youth comply with the probation order and report to the probation officer
- notify of any change of address, employment, education or training
- remain within an area specified by the probation order
- make reasonable efforts to obtain and maintain suitable employment
- attend school or other appropriate training
- reside with a parent or other such adult considered appropriate
- reside in a place specified by the probation order
- comply with any other reasonable conditions set out in the probation order
While probation was used extensively as a disposition either alone or in combination with other dispositions, there was no relationship between having a probation order and the index of child-like or adult-like characteristics.\textsuperscript{23}

Table 37

<table>
<thead>
<tr>
<th>Number of child-like factors</th>
<th>Probation</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>no</td>
<td>yes</td>
<td>Total</td>
</tr>
<tr>
<td>adult or 0 child</td>
<td>4</td>
<td>15</td>
<td>19</td>
</tr>
<tr>
<td>1-2 child</td>
<td>9</td>
<td>23</td>
<td>32</td>
</tr>
<tr>
<td>3 or more child</td>
<td>9</td>
<td>25</td>
<td>34</td>
</tr>
<tr>
<td>Total</td>
<td>22</td>
<td>63</td>
<td>85</td>
</tr>
</tbody>
</table>

Not significant.

**Probation conditions:**

Perhaps then, the use of probation as a disposition is not related to age, but rather the conditions placed within the probation order may relate in some way to age related concerns.

Table 38 shows the different kinds of probation conditions placed on youth in this sample, by the frequency of use of each condition.

\textsuperscript{23} Analyses were completed on all of the different combinations of dispositions used: custody only, custody, probation and intermediate sanction in one sentence, custody and probation only, custody and an intermediate sanction only, probation only, probation and intermediate sanction only, intermediate sanction only. None of these combinations related to either the child-like index or the variable on chronological age group.
Table 38 - Probation conditions for young offenders in this sample

<table>
<thead>
<tr>
<th>Probation Condition</th>
<th>number/64</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>report to probation officer</td>
<td>41</td>
<td>64%</td>
</tr>
<tr>
<td>no keeping company of...co-accused...victim(s), etc.</td>
<td>32</td>
<td>50%</td>
</tr>
<tr>
<td>not in possession of ... weapons ... non-medically prescribed drugs</td>
<td>21</td>
<td>33%</td>
</tr>
<tr>
<td>obtain counseling</td>
<td>21</td>
<td>33%</td>
</tr>
<tr>
<td>attend school/employment</td>
<td>17</td>
<td>27%</td>
</tr>
<tr>
<td>boundary restrictions</td>
<td>13</td>
<td>20%</td>
</tr>
<tr>
<td>obey written/house rules</td>
<td>12</td>
<td>19%</td>
</tr>
<tr>
<td>curfew</td>
<td>9</td>
<td>14%</td>
</tr>
<tr>
<td>write a letter</td>
<td>3</td>
<td>5%</td>
</tr>
<tr>
<td>victim reconciliation</td>
<td>2</td>
<td>3%</td>
</tr>
</tbody>
</table>

An analysis of these probation conditions in relation to age shows that there are no significant relationships, with the exception of boundary restrictions\(^{24}\).

Table 39

<table>
<thead>
<tr>
<th>number of child-like factors</th>
<th>boundary restrict</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>adult or 0</td>
<td>no</td>
<td>14</td>
</tr>
<tr>
<td>child</td>
<td>yes</td>
<td>93.3%</td>
</tr>
<tr>
<td>1-2 child</td>
<td>no</td>
<td>21</td>
</tr>
<tr>
<td></td>
<td>yes</td>
<td>91.3%</td>
</tr>
<tr>
<td>3 or more child</td>
<td>no</td>
<td>16</td>
</tr>
<tr>
<td></td>
<td>yes</td>
<td>61.5%</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>51</td>
</tr>
</tbody>
</table>

Chi-square=8.934, \(df=2, p=.011\), 2 expected values less than 5(3.05), 0-2 (pooled) vs. 3 or more, Fisher's exact test \(p<.05\).

\(^{24}\) Boundary restrictions are orders of the court whereby individuals are barred from entering into a particular geographical area or being within a certain distance of that area in order to avert further potential offending, or to avert contact with victim(s) and/or co-accused(s).
Table 39 shows that boundary conditions are more likely to be placed on youth with more ‘child-like’ characteristics. With up to 2 ‘child-like’ characteristics, few boundary conditions are placed on youth, but with 3 or more ‘child-like’ attributes, youth are at least 4 times more likely than those with up to 2 child-like factors to be given boundary conditions as part of their probation order.

**Age group**

There was also a relationship between the age group of a young offender and the likelihood of having a boundary condition placed upon their probation order. As Table 40 demonstrates, older youths were less likely to be given a boundary condition.

**Table 40**

<table>
<thead>
<tr>
<th>age group</th>
<th>boundary restrictions?</th>
<th>no</th>
<th>yes</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>12-13</td>
<td></td>
<td>2</td>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td></td>
<td>40.0%</td>
<td>60.0%</td>
<td>100.0%</td>
</tr>
<tr>
<td>14-15</td>
<td></td>
<td>14</td>
<td>6</td>
<td>20</td>
</tr>
<tr>
<td></td>
<td></td>
<td>70.0%</td>
<td>30.0%</td>
<td>100.0%</td>
</tr>
<tr>
<td>16+</td>
<td></td>
<td>34</td>
<td>4</td>
<td>38</td>
</tr>
<tr>
<td></td>
<td></td>
<td>89.5%</td>
<td>10.5%</td>
<td>100.0%</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>50</td>
<td>13</td>
<td>63</td>
</tr>
<tr>
<td></td>
<td></td>
<td>79.4%</td>
<td>20.6%</td>
<td>100.0%</td>
</tr>
</tbody>
</table>

Pooling 12-15 year olds vs. 16+, Fisher’s exact test $p=.029$

Thus, ‘younger’ youths may be seen by the court as needing restrictions on the areas in which they can enter in order to avert the potential for further criminality.
Number of conditions placed on youth by age

Finally, an analysis of the number of conditions placed upon youth was conducted to see if more probation conditions were placed upon younger offenders in order to more intensively supervise their behaviour. The results show that neither the presence of child-like or adult-like factors, nor the chronological age group to which the youth belonged were significantly related to the number of probation conditions placed on the order (Tables 41 and 42).

Table 41

<table>
<thead>
<tr>
<th>Relationship between child-like factors and the number of probation conditions</th>
<th>0-1 cond</th>
<th>2-3 cond</th>
<th>4+ cond</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>number of child-like factors</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>adult or 0 child</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>15</td>
</tr>
<tr>
<td>1-2 child</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1-2 child</td>
<td>6</td>
<td>10</td>
<td>7</td>
<td>23</td>
</tr>
<tr>
<td>3 or more child</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3 or more child</td>
<td>5</td>
<td>9</td>
<td>12</td>
<td>26</td>
</tr>
<tr>
<td>Total</td>
<td>15</td>
<td>24</td>
<td>25</td>
<td>64</td>
</tr>
<tr>
<td>Not significant.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Section IV - Pre-disposition reports

Pre-disposition reports are used in both youth and adult courts to provide a judge with contextual information about the offender and the circumstances of the offence.

Under Section 24(1) of the YOA, a pre-disposition report must be ordered if a custodial sentence is being considered unless, with the consent of the prosecutor and the young person or his counsel, it is determined that the report is unnecessary or not in the best interests of the youth to require one (Section 24(3) YOA). Generally, the pre-disposition report allows the judge to assess mitigating circumstances as an element of fairness in passing sentence (Sarkesian 1989; Nadin-Davis 1982). Section 14 of the YOA sets out the categories of information that a Pre-Disposition Report shall include:

- the results of an interview with the young person, if possible their parents, and potentially members of the young person’s extended family
- the results of an interview with the victim in the case
- information on the age, maturity and character of the young person and his willingness to make amends
- plans put forward by the young person to change his conduct
• the history of previous findings of delinquency
• the history of the use of alternative measures with the young person
• the availability of appropriate community services
• the relationship between the young person and the young person’s parents and the degree of control and influence the family has on the young person
• school attendance and performance and employment record
• any other appropriate information

There has been a great deal of research which has examined the use of pre-disposition reports in terms of their effects on sentencing outcomes, the general contents and quality of the reports, and the social construction of the offender in the report. It has been shown that judges rely quite heavily on pre-disposition reports when they are available (Hanscom 1988).

Studies show that the pre-sentence investigation is crucial in assessing the juvenile’s character for court personnel (Sanborn 1996), that judges are confident with the accuracy of reports (Hanscom 1988), and that judges rate the pre-disposition report as a primary source of information (Holsinger and Latessa 1999, Hanscom 1988). In Hanscom’s study, judges felt that pre-disposition reports were not biased in that they did not form the majority of either Crown or Defense counsel’s cases (Hanscom 1988). However, whether or not the pre-disposition report benefits or hinders a young offender’s case is an area of contention. A study by Milne, Linden and Kueneman (1992) showed that defense lawyers were uncertain as to whether or not the contents of the pre-disposition reports helped or hindered their clients’ case. The authors note that certain lawyers expressed concern with the inadmissibility of information contained within pre-disposition reports “...in the form of inflammatory statements and hearsay reflecting the

---

25 The recently proposed Act to replace the YOA called the Youth Criminal Justice Act has added two other categories of information that shall be included in a pre-sentence report (1) the recommendations resulting from any conference (2) any information on alternatives to custody (Bill C-3 (1999) Section 39). In
bias of the particular probation officer” (Milne, Linden and Kueneman 1992: 340). Other defense lawyers thought pre-disposition reports to be useful:

...I think that the reports are useful to the defence. The judge gets the idea that he's dealing with a person who is not just a case, but a human being (Milne, Linden and Kueneman 1992: 341).

Despite the likelihood that the pre-disposition reports will facilitate a more individualized approach to sentencing, the potential still remains for an unfair or biased construction of the offender within these reports. In her study of the construction of gender in pre-disposition reports, Sarkesian notes;

...probation officers do engage in defining the lives of male and female offenders on different dimensions. They use selective investigation and reporting strategies which shape images and perceptions of male and female offenders. ...The principle of individualized justice facilitates this subtle and overt form of discrimination by suggesting that the individual differences provide for mitigating circumstances and therefore, make for better judgements. Probation officers and other decision makers must be aware of these very subtle forms of constructing and reproducing further inequalities (Sarkesian 1989: 46,47).

Thus, the contents of Pre-Disposition Reports have important implications for the outcome of the case since, in the majority of cases, judges concur with the recommendations made by probation officers26 (Markwart 1992; Gelsthorpe and Raynor 1995). Moreover, as Hanscom notes, while judges assume that the information in Pre-Disposition Reports is generally accurate, the information in the report is rarely scrutinized. The cross-examination of probation officers on the contents of reports rarely

---

26 As of 1995, section 14(d) of the YOA explicitly states that the provincial director (probation officer) can include a recommendation for sentence in the pre-disposition report, although the judge is under no obligation to follow this recommendation.
occurs, and the presence of a probation officer at the sentencing hearing is also quite infrequent (Hanscom 1988).

Pre-disposition reports were used in about 21% of the cases in this sample. The results indicate that there were no relationships between the principal charge and the use of a pre-disposition reports in the sentencing hearing (Table 43). This may be explained by the variation in each category of ‘principal’ charge as to the seriousness of the offence.

Table 43

<table>
<thead>
<tr>
<th>principal charge</th>
<th>pre-disposition report used?</th>
<th>no</th>
<th>yes</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>violence</td>
<td></td>
<td>20</td>
<td>6</td>
<td>26</td>
</tr>
<tr>
<td></td>
<td></td>
<td>76.9%</td>
<td>23.1%</td>
<td>100.0%</td>
</tr>
<tr>
<td>drugs</td>
<td></td>
<td>5</td>
<td>3</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td></td>
<td>62.5%</td>
<td>37.5%</td>
<td>100.0%</td>
</tr>
<tr>
<td>b&amp;ie</td>
<td></td>
<td>3</td>
<td>3</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td></td>
<td>50.0%</td>
<td>50.0%</td>
<td>100.0%</td>
</tr>
<tr>
<td>property&amp;occo</td>
<td></td>
<td>40</td>
<td>6</td>
<td>46</td>
</tr>
<tr>
<td></td>
<td></td>
<td>87.0%</td>
<td>13.0%</td>
<td>100.0%</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>68</td>
<td>18</td>
<td>96</td>
</tr>
<tr>
<td></td>
<td></td>
<td>79.1%</td>
<td>20.9%</td>
<td>100.0%</td>
</tr>
</tbody>
</table>

Not significant.

There was however, a statistically significant relationship between the use of a pre-disposition report in a case and the young offender having a prior record.
Table 44

Relationship between a prior record in the case and the case having a pre-disposition report

<table>
<thead>
<tr>
<th>is there a PDR in court case?</th>
<th>no</th>
<th>yes</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>is there a prior record?</td>
<td>no</td>
<td>38</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>yes</td>
<td>90.5%</td>
<td>9.5%</td>
</tr>
<tr>
<td></td>
<td>yes</td>
<td>26</td>
<td>14</td>
</tr>
<tr>
<td>Total</td>
<td>64</td>
<td>18</td>
<td>82</td>
</tr>
</tbody>
</table>

Fisher’s exact test $p=.007$

Youths with a prior record were about 3 times more likely to have a pre-disposition reports in their sentencing hearing. This may, in part, be accounted for by the fact that youths with prior records were more likely to be given a custodial sentence, and that those with prior records are more likely to already have a pre-disposition report previously written on their case history for the use of the court.

Of the 18 cases where a pre-disposition report was used in determining the sentence, I was able to access 13 of these reports. An examination of these reports indicates that the court ‘agreed’ or ‘mostly agreed’ with the recommendation of the probation officer in over half (7/13) of the cases. In 2 cases, the court partially followed the recommendation (e.g. used custody but not the length suggested in the report). In 4 cases the recommendation was not followed at all.
Child-like/Adult-like language used in Pre-Disposition Reports:

An assessment was made as to the degree to which pre-disposition reports construct the 'youthfulness' of an offender in these reports and whether or not this relates to the outcome of the case (Table 45).
Table 45 - Sample of Pre-Disposition Reports showing references to age and family situation

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>2</td>
<td>2 0</td>
<td>Separated</td>
<td></td>
<td>-</td>
<td>-</td>
<td>assault</td>
<td>23 days s/c; 12 m probat</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>2 1</td>
<td>brief marriage</td>
<td>father-abusive</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>30 days s/c</td>
</tr>
<tr>
<td>0</td>
<td>4</td>
<td>0 1</td>
<td>no difficulties w/pregnancy &amp; delivery; met all “milestones”</td>
<td>Married</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>23 days o/c; 12 m probat</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>1 0</td>
<td>youth was result of rape of mother @ 18 yrs old at a drug party</td>
<td>-</td>
<td>no contact</td>
<td>on Mother’s Allowance</td>
<td>-</td>
<td>30 days o/c; probatio compns</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0 0</td>
<td>-youth born a healthy baby-reached all “milestones” early</td>
<td>never married</td>
<td>-</td>
<td>mother unemployed</td>
<td>possess under</td>
<td>12 m probat 25 hrs ca</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>0 0</td>
<td>-unexpected birth, no complications w/pregnancy</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>40 hours ca; 18 mths probat</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0 0</td>
<td>birth &amp; early develop-normal - youth in car accident at 9 - suffered head injuries</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>30 days o/c; 18 m probat</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>1 0</td>
<td>-lives with stepmother; real mother whereabouts unknown</td>
<td>step mother seeking employment w/ no success</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>56 days o/c; 12 m probat</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>0 0</td>
<td>-never married to father</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>escape custody x2</td>
<td>1 day s/c</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>0 0</td>
<td>-never married to father</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>traffick x3</td>
<td>2 months s/c; 12 m probat</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>0 0</td>
<td>-parents - divorced</td>
<td>-</td>
<td>-</td>
<td>mother - family benefits</td>
<td>-</td>
<td>-2 weeks o/c</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>1 0</td>
<td>-parents - separated - mom has new partner</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>15 days o/c; 16 m probat</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0 0</td>
<td>- &quot;awful marriage&quot; - now separated</td>
<td>-</td>
<td>-</td>
<td>robbery x4</td>
<td>-</td>
<td>30 days s/c; 6 m o/c; 18 m probat</td>
</tr>
</tbody>
</table>

NOTE: Youth - positive "youth listens to music" "plays video games" "attends school regularly" Youth - negative "youth is out of control" "late for school/truant" "goes to parties all night" "talks on phone all night" Youth/Adult - "youth is influenced/recruited by older peer group" "easily influenced by older friends" Adult - "was employed F/T as a general labourer" "presents herself as older"
It is difficult to make any generalizations about the severity of sanctions based upon 'child-like' or 'adult-like' constructions of youths in these reports because all cases, except two, received custody. The two cases that received sanctions other than custody do not appear to differ substantially from the rest in terms of the way the youth was described.

What can be said about this small sample of reports is that the qualitative descriptions of these youths vary among reports. In a few of the pre-disposition reports, there is a more consistent characterization of the 'youthfulness' of the offender in either positive or negative terms (eg. attends school regularly, listens to parents, obeys curfew - all constructing 'youthfulness' in positive terms). However, in other reports, descriptions are not limited to constructing 'youthfulness' as a mitigating factor which warrants a less harsh approach, but 'youthfulness' is also constructed as an aggravating factor, or a justification for increased social control. For example, in one report, the probation officer cites the arresting police officer's description of the youth at arrest;

...Constable 'X' followed him into the washroom and reported that the two of them were in the only cubicle, there was heroin on a piece of foil and 'Y' held a cigarette and a rolled up five dollar bill. Both were smoking heroin. As Constable 'X' entered the washroom 'Y' flushed everything down the toilet. 'Y' then started crying and calling for his mother which Constable 'X' explained is 'Y's usual role. He pretends to be innocent and can be a convincing actor but according to Officer 'X' he is not innocent and will reoffend in a way that is more beneficial to him....

In the same report a few paragraphs later, the writer notes that...

"Y' was frequently truant which did lead to his bail being revoked. 'Y' does say that he needs an education to be able to get a job and does like computers, math and design technology. ...'Y' was able to verbalize what he needs to do to stay out of trouble, "get a job, go to school and listen to the rules". ...he enjoys going to Canada's Wonderland, lifting weights and playing soccer and basketball with his cousins.
In this report, as with others, there are constructions of the youth as being on the one hand relatively responsible and being a typical young person (e.g. knowing what it takes to get a job in computers, going to Canada’s Wonderland and playing with cousins) and on the other hand, needing supervision (e.g. truant, pretending to be innocent in the officer’s description etc.). From this small sample of reports then, it is unclear what the overall construction of the youth is, or how the varied information on youthfulness does or does not impress upon the courts when determining sentencing decisions.

The other interesting point that comes out of these reports does not relate to age, but rather to the family situation. In this sample of reports there were a number of categories or themes devoted to the mother’s pregnancy, the parent’s marital situation, and a recurring theme of paternal abandonment. Presumably, the reason this kind of information is included is due to underlying ‘theories’ of why the youth was involved in criminal activity and the amount of supervision the family can provide. For example, in just under half of the reports (5/13), there was information related to early childhood development and even pre-natal information. In one case, as noted in Table 45, the information presented on the youth pertained to the nature of the youth’s conception. “...Mrs. ‘X’ was pregnant with ‘Y’ when she was about 18 years old. Her pregnancy was a result of a rape that occurred during a drug party that she was attending.”

In 10 out of 13 cases, information was presented regarding the marital situation of parents. Of these only 2 of the parental unions were characterized as stable relationships. Finally, in 9 of 13 cases, information on the father was presented in the report. In all of these cases there was a theme of paternal abandonment or absence in the youth’s life. In one case the father was said to have died under “mysterious circumstances”. In most,
there was no contact from the father, and in many there were characterizations of the father being domineering, abusive, or having a problem with alcohol.

**Conclusions**

In the context of the *Young Offenders Act* which suggests that age and maturity should be factors which are taken into account in sentencing decisions, this chapter sought out to examine if, and how, age and the 'youthfulness' of the offender related to the outcome of a sample of sentencing cases. In addressing this question, this study reveals that there is no substantial relationship between age and the use of custody. The one exception to this is—in the few cases where there was an adult co-accused, the courts were more likely to impose a custodial sanction. However, having a case with an adult co-accused did not independently predict the likelihood of custody. Instead, other legal variables—prior record and pre-trial detention—were independently related to the likelihood of cases receiving custody. To some extent as well, the perceived ethnicity of the youth predicted the use of custody.

This is consistent with the analysis of data across Canada and in the provinces of Ontario, Quebec and British Columbia. With the exception of robbery, which encompasses a wide range of offences that vary dramatically in their 'seriousness', there was no clear relationship between age groups and the use of custody even when controlling for the presence or absence of a previous record.

Moreover, principles of the *Young Offenders Act* and goals of sentencing were rarely raised in the observed court hearings, and when they were raised, they were in reference to public protection and the public interest. In very few cases was there explicit
mention of the principles relating to young people and their special circumstances.

Similarly, in the cases where custody was being considered, there were few cases in which there was mention of the guidelines in Section 24 of the Young Offenders Act which sets out when it is appropriate to use a custodial sentence for a young person.

Even the use of probation conditions did not appear to relate to the age group of the youthful offender. Neither the number of conditions nor the kinds of conditions placed on youth related to their age, except for ‘boundary conditions.’ Boundary conditions appeared more frequently on cases which dealt with younger youth (12 and 13 years old) and youths who had characteristics that made them appear more ‘child-like’. Thus, it may be the case that the use of this condition is an attempt to protect the younger age group from going into certain areas and ‘getting into more trouble’.

Finally, the analysis of pre-disposition reports in this sample showed that while these youths were described in both ‘youthful’ and ‘adult-like’ terms, this small sample of reports did not reveal a pattern one way or another in terms of the likelihood of receiving custody (since all but 2 cases received custody), and in terms of the report consistently characterizing the offender in either ‘youthful’ terms of ‘adult-like’ terms. Instead, what this analysis revealed is that constructions of youth based upon age vary from report to report, but that in general, descriptions portray youthfulness as a mitigating factor (attends school) or an aggravating factor (is truant) in the same report. In the final analysis it appears that, as with the chapter on bail hearings, decisions made in youth court sentencing hearings are predominantly governed by legal variables.
Appendix A - Sentencing coding sheet

(a) Location & time
address ____________________________
courtroom# _______________________
date__/___/________
d an pm ___________________________
judges __________________________
(b) Case information
sentencing hearing trial
docket/file# _______________________
name __________________________
age ____________ perceptual age ______
d.o.b. (dd/mm/yy) __________________
phase [ ] phase [ ] not applicable
sex: [ ] m [ ] f
in custody? y n unclear
S/C [ ] O/C
(c) Ethnicity:
[ ] Caucasian
[ ] Black
[ ] Asian/Oriental
[ ] South Asian
[ ] Aboriginal
[ ] Hispanic
[ ] Other
[ ] Unknown
name of Crown [ ] Provincial [ ] Federal
(d) Lawyer:
lawyer’s name _______________________
is present with youth?
[ ] Yes [ ] No
who represents youth @ hearing?
[ ] Duty counsel
[ ] Privately retained name?
[ ] Student at law (if different from above)
[ ] Unavailable
[ ] Nobody
(d) Current Charges?
what is/are the current charge(s)/allegation(s)
number of charges read in: __________
[ ] None
[ ] Not mentioned
[ ] Yes, is one or more of the prior(s)
[ ] Related to current charge
[ ] Unrelated to current
[ ] Unstated
[ ] Recent (within the last year)
[ ] Not recent (over one year ago)
[ ] Unstated
[ ] Includes violence
[ ] Does not include violence
[ ] Unstated
(e) Co-accused
do one or more offences involve at least one other offender?
[ ] Yes
[ ] No
if yes, is co-accused(s) present?
[ ] Yes
[ ] No
[ ] Yes: also in custody
[ ] Unstated
is one or more co-accused an adult?
no yes no mention
what role does the v.o. being sentenced today have in the offence?
☐ equal to other offender(s)
☐ greater than other offender(s)
☐ less than other offender(s)
☐ not mentioned

(if family:
are parent(s) guardian(s) present in court?
☐ yes ☐ no
☐ apparently (kid sitting w/ adults)
☐ not obvious/unclear
☐ v.o. & parent(s) not waiting in courtroom

please charge/allegation(s)

1. g ng
2. g ng
3. g ng

if no: why?
☐ parent(s) at work
☐ in hospital
☐ didn't want to come
☐ no contact with v.o.
☐ not mentioned

is v.o. sitting beside parent(s)?
☐ yes
☐ no
☐ v.o. is in custody
☐ not obvious
☐ unknown
☐ not applicable

parent(s) involved in v.o. 's life?
☐ yes
☐ no
☐ not mentioned
☐ not applicable

do they wish to be involved in helping youth with offence?
☐ yes
☐ no
☐ not mentioned
☐ not applicable

are others present in court?
☐ yes
☐ no
☐ unclear
if yes: who?
☐ sibling(s)
☐ grandparent(s)
☐ other family
☐ other(s)
☐ C.A.S.
☐ friend(s)

are parents able to supervise v.o. (i.e. work FT)
☐ yes
☐ no
☐ not mentioned
unlikely
☐ not applicable

is parent(s)/guardian(s) occupation mentioned?
☐ no
☐ yes:

is/are parent(s)/guardian(s) on govt assistance?
yes Family Benefits(Mother's Allow)
Disability (Workman’s Comp)
Social Assistance (Welfare)
U.I.C.

no
not mentioned
(g) Living arrangements:
does youth live w/parent(s)/guardian(s)?
☐ yes
☐ no
☐ no mention
☐ not applicable
if no: where does youth live?
☐ with other relative
☐ at friend’s house
☐ on the street
☐ at child welfare agency
☐ with girlfriend/boyfriend/CLaw
(b) youth’s appearance:
how is youth dressed for court?
☐ dressed up (suit, dress pants, dress shirt)
☐ ‘neat’ middle class (jeans, dress shirt)
☐ dressed down (jeans, t-shirt, sweatshirt)
other:
are there other ‘non-child’ or ‘child’ signs about the youth?
☐ no
☐ yes:
piercings
gaistee
moustache
peach fuzz
youth is a parent
other:
are there any signs about youth that show disinterest/disrespect in proceedings?
☐ no
☐ yes:

result
1. g ng
2. g ng
3. g ng

how interested is youth in proceedings? (mark 1+)
☐ very interested
☐ moderately interested
☐ disinterested/looking around court
☐ defiant looking
☐ laughing (w/ friends)/aloof
☐ school:
in school?
☐ no
☐ yes
name of school
last grade completed

☐ not mentioned
☐ most of the time
☐ sometimes
☐ once in a while
☐ is a student - but not presently registered
☐ kicked out

how well is y.o. doing in school?
☐ excellent marks
☐ very good
☐ good
☐ average
☐ passing all courses
☐ failing
☐ not mentioned
☐ not applicable

is youth involved in extra-curricular activities?
☐ yes
☐ no
☐ not mentioned
   unlikely
if yes: what?

is there other info. presented from school?  
☐ yes  
☐ no  
(ie. comments - teacher/principal/counsellor)  
Does youth have a learning disability?  
yes  no  no mention  
Is v.o. on/or ever been on welfare?  
yes  presently on welfare  in the past  
no  
not mentioned  

does youth work?  
yes  where?  
   F/T  P/T  
no  
not mentioned  

comments:

sentence:
1.  
2.  
3.  
☐ charges withdrawn

(j)sentencing hearing:  
is there a PDR in this case?  
☐ yes  
☐ no  
if yes: is info presented from it?  
☐ yes  
☐ no  
if yes: what
Is there a joint submission from Crown and Defense?  
☐ yes  
☐ no  
if yes: what is it?  
if no: what info is presented by Defense?  
does Defence give a case plan for v.o.?  
☐ yes  
☐ no  
specify:  
it no to joint submission. Crown's position?

does judge ask youth to make any further comments before sentencing?  
☐ yes  
☐ no  

(k)sentence:  
if joint submission, does judge agree?  
☐ yes  
☐ no
judge's comments @ sentencing:

final submissions
Crown
Defense

if trial: other information from witnesses

witness #1

witness #2

sentence
☐ secure custody
☐ open custody
☐ probation
☐ community service
☐ compensation
☐ restitution
☐ conditional discharge
☐ absolute discharge

conditions:
☐ keep peace - be of good behaviour
☐ obey rules of house
☐ curfew: _______ to _______
☐ attend school everyday
☐ not in possession of non-medically
pres. Narcotic
☐ not in possession of weapon
☐ no keeping company of:
  ☐ co-accused(s)
  ☐ victim(s)
  ☐ anyone known to have crim rec

☐ reside with _______
☐ report to prob officer as required
☐ obtain counselling as specified
Appendix B - Inter-rater reliability

<table>
<thead>
<tr>
<th>level of discrepancy in observation between raters</th>
<th>#</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>total number of observations made together</td>
<td>569</td>
<td>100%</td>
</tr>
<tr>
<td>total agreement between raters</td>
<td>508</td>
<td>89.3%</td>
</tr>
<tr>
<td>slight differences between raters (e.g. not clear vs. yes, not mentioned vs. yes, perceived age within 2 years)</td>
<td>53</td>
<td>9.3%</td>
</tr>
<tr>
<td>total disagreement between raters (perceived age more than 2 years etc.)</td>
<td>7</td>
<td>1.2%</td>
</tr>
</tbody>
</table>
# Appendix C - Breakdown of Indices

<table>
<thead>
<tr>
<th>Index Name</th>
<th>Variables Included</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Legal</strong></td>
<td>- youth was held in detention before the sentencing hearing (+)</td>
</tr>
<tr>
<td></td>
<td>- principal charge is either 'violence' or 'drugs' (+)</td>
</tr>
<tr>
<td></td>
<td>- youth has a prior record (+)</td>
</tr>
<tr>
<td></td>
<td>- 3 or more charges were read into at hearing (+)</td>
</tr>
<tr>
<td></td>
<td>- prior record has offences that are related to current offence (+)</td>
</tr>
<tr>
<td></td>
<td>- prior record has offences that are recent (committed within the last year) (+)</td>
</tr>
<tr>
<td></td>
<td>- prior offences include violence (+)</td>
</tr>
<tr>
<td></td>
<td>- there is a co-accused in the case (+)</td>
</tr>
<tr>
<td></td>
<td>- the co-accused is present in court on day of -or also in custody (+)</td>
</tr>
<tr>
<td></td>
<td>- the role of the youth in this offence is said to be greater than or equal to co-accused (+)</td>
</tr>
<tr>
<td><strong>Child-like/Adult-like</strong></td>
<td>- the real age of the youth is 12 through 15 years (+)</td>
</tr>
<tr>
<td></td>
<td>- the perceived age of the youth is 15 years or less (+)</td>
</tr>
<tr>
<td></td>
<td>- the youth lives with their parent (+)</td>
</tr>
<tr>
<td></td>
<td>- the youth is in school (+)</td>
</tr>
<tr>
<td></td>
<td>- the youth is involved in extra-curricular activities (+)</td>
</tr>
<tr>
<td></td>
<td>- the youth exhibits other 'child-like' signs (ie. is crying on the stand) (+)</td>
</tr>
<tr>
<td></td>
<td>- there is a co-accused who is in adult in the case (-)</td>
</tr>
<tr>
<td></td>
<td>- the youth has facial hair ('males only') (-)</td>
</tr>
<tr>
<td></td>
<td>- the youth has been or is presently on welfare (-)</td>
</tr>
<tr>
<td></td>
<td>- the youth exhibits one+ adult-like signs (ie. is a parent) (-)</td>
</tr>
<tr>
<td></td>
<td>- the youth is employed full-time(-)</td>
</tr>
<tr>
<td><strong>Support</strong></td>
<td>- a parent/guardian is present at court, or the court hears why the parent/guardian could not come, and this is for legitimate reasons (ie. at work) (+)</td>
</tr>
<tr>
<td></td>
<td>- the parent/guardian is said to be involved in youth's life (+)</td>
</tr>
<tr>
<td></td>
<td>- the parent/guardian is able to supervise this youth (+)</td>
</tr>
<tr>
<td></td>
<td>- defense provides a concrete case plan for the youth (+)</td>
</tr>
</tbody>
</table>

+ means that this factor counted towards the overall index  
- denotes that this factor dettracted from the overall index
### Appendix D - Detailed information on cases where reference was made to principles and/or goals of sentencing

| ID# 2 | Crown recommends 2 years probation for the purposes of rehabilitation and individual deterrence  
offence - theft under: obstruct peace officer; carry concealed weapon  
sentence - 7 days S/C; 24 months probation w/ conditions - rt in 7 days; no weapons-knives; associations with criminals; obey written rules, school/empl | rehabilitation - C  
specific deterrence - C |
| ID# 17 | Judge - "This order of disposition has to help you say "no" to crime" - "not an order to punish you" - "personal deterrence is important" - "you're not alone-your mother is with you" - but you have committed 3 serious offences—you need an order to think about your choices/friends  
offence: traffic narcotic; possession for the purposes of trafficking; (crime proceeds -dropped)  
sentence - s/c - 2 months; probation 2 months—attend school, reside where directed, obey written rules of residence and report/2 weeks | specific deterrence - J+ |
| ID# 19 | Defense - "18 months is unduly harsh and not conducive to the YOA principle of rehabilitation" - "consider general deterrence and specific deterrence"  
Judge: what about safety of the public? These are the most serious & dangerous offences that young people can commit—he panicked and hit the gas  
Defence: "He has learned a very serious lesson" - "suggest a short period of open custody - and consider an extended period of probation if you're concerned with specific deterrence"  
Crown(cross examination) - rehabilitation is important but social denunciation is just as important in this case  
offence: dangerous driving; poss over 5000x3; theft over 5000 x3  
sentence: 56 days O/C and 12 months probation—report to p. officer, attend program related to driving offences; no b/w driving/learning to drive |
| ID# 22 | Crown - "seeking something shorter and sharper" - "s3(b) society must be afforded necessary protection (both up and general)" "general deterrence is a valid principle to apply" - "general deterrence will not be served by probation"  
offence: trafficking mj (06 grams)  
sentence: 23 days O/C; 12 months probation |
| ID# 19 | Defense (to judge) - "all of his problems occurred in a very short period of time as a result of a breakdown in the family" - "consider the principles of the YOA" - "he has already spent 3 months at West Detention Centre" "he panicked when approached by the officers and dove towards the officers by mistake" "this accused came forward and took responsibility for his actions...this accused is more mature than co-accused and is more reflective of family problems"  
rehabilitation - D+ C+  
specific deterrence - D  
public protection - J+  
denunciation - C+  
general YOA principles - D+ |
| ID# 22 | Defense - "any sentencing principles can be met by the maximum hours of community service" - "follow YOA principles to balance needs of society with needs of the individual"  
rehabilitation - public protection - C+  
specific deterrence - C+  
general YOA principles - D+ |
| ID# 34 | Judge - "this is not against the public interest"  
offence: theft under 5000; poss under 5000  
sentence: absolute discharge |
| ID# 39 | Defense - wants o/c - for rehabilitation and s/c for general deterrence - "the more time spent with bad influences the more toxic he becomes - want s/c - short sharp to get message and longer o/c"  
Judge - "I want a long term s/c -but this is the last chance for rehabilitation"  
offence: robbery x4; dangerous driving; weapons dang. FTC with bail  
sentence: 30 days s/c; 6 months o/c 18 months probation |
| ID# 39 | Judge - crime prevention - long term protection  
Judge - can project society via rehabilitation  
rehabilitation - D, J  
general deterrence - D  
crime prevention - J  
public protection - J |
| ID# 43 | Crown - wants probation "not in the public interest to let him off--discharge" - "he was supposed to be on house arrest & was out in stolen vehicle" (Defense wants probation via conditional discharge)  
offence: possession over $5000; FTC with recognition  
sentence: conditional discharge through 6 months probation |
| Public interest - C |
**ID#46 - Crown** - "there's a reason that society requires denunciation and general deterrence – because kids under age 14 should not be allowed to consent" “no chance for rehabilitation because of surrounding circumstances”  
**Offence**: sexual assault  
**Sentence**: 40 days O/C; probation 12 months (joint sub)

**ID#47 - Judge** - "has spent 7 days pre-trial custody; no prior record; 7 days in pre-trial custody and 1 day O/C additionally will serve gen and spec deterrence.  
denunciation and protection of society."
**Offence**: breach of probation - curfew; fail to comply w/ recog.
**Sentence**: (7 days detention) 1 day O/C

**ID#48 - Judge** - "there's very little jurisprudence that gives the court guidance as to when the court can exercise a conditional discharge" (Judge makes reference to principles relating to adult offenders (wrt to conditional discharge and probation) "... in youth court, being placed on probation doesn't carry the same stigma as it does in adult court. Parliament has given this new option (of conditional discharge) to get us more alternatives. Taking into consideration the purposes and principles of the YOA the fact that you failed to attend court because you slept in doesn't give me reasons to use a conditional discharge - should get probation"

**ID#49 - Defense** - "consider a discharge - not against the public interest"
Judge - "OK absolute discharge, but given prevalence of auto theft - not in public interest to give discharge"  
"damage etc... undertaking to pay restitution - your dad says boys will be boys - I say grow up"
**Offence**: auto theft $300 damage
**Sentence**: $300 restitution

**ID#50 - Crown** - general reference to general deterrence  
Judge: "Crown wants me to put you in Metro West which is overcrowded and full of y.o. s with long records. Your lawyers says the agony of you having to come to court4/5 times over the course of a year is enough to convince you .. “Do I take a chance on X and go with a positive PDR ?” “I think I'll take a chance but there'll still be some payback - I don't want those on the streets of Scarborough to think that the courts don't care restores owners (general det). But - I will consider factors that you weren't the main player & the agony that you have gone through where you were told last fall that Crown was seeking 6-8 months S/C & I think what you've achieved this past year has been motivated by this change. Your attendance at school was good. You are now an adult. If your conditions are breached you'll be in adult court and that's a different ball game. Community service is a payback to society to show you're remorseful and an alternative to incarceration. I hope you live with those terms and that we don't see you back here because if we do then you'll be one of the consequences"  
**Offence**: robbery
**Sentence**: 100 hours CSO
**ID# 77 - Judge:** general reference to young person's need to be held responsible for their behaviour; can protect society through rehabilitation.

- Offence: theft under, FTC with bail (branched currency) FTA
- Sentence: 35 days PTD; 18 months probation; 40 hours CSO

**ID# 79 - Crown:** "y.o. demonstrated no remorse at time of arrest - fail to appear for initial interview, probation officer said he had difficulty in O/C & his response to S/C was more favourable" "general deterrence is important"

Deterrence - "he's had a taste of the justice system and has stayed out of trouble for over a year now - this is an indication that he's been rehabilitated & general deterrence is a minor principle in this case - the governing principles are specific deterrence and rehabilitation as held in case law (judge said that these are the most important principles)

Judge: "you demanded money from a teller and said 'or I'll shoot you'. General deterrence is a principle that must be considered but it's not the most important..."

"he's an atypical offender that is seen before the court. The problem of that though is that he's said to be a good, friendly, happy person, however he has accumulated a significant criminal record. I must also take into consideration the date of this offence - 2 factors to consider:

1) the interim time - just because apprehended later doesn't mean they don't suffer the consequences. PDR page 4 says he stopped attending school (despite probation order to attend school) pg.5 PDR says he's been tardy following C.S. orders. Pg6 he failed to attend initial interview - this all indicates a concern with rehabilitation. He has entered a guilty plea and has straightened out other aspects of his life. O/C won't serve general or specific deterrence so need S/C..."

- Offence: robbery
- Sentence: 16 days S/C; 1 year probation

**ID# 82 - Crown:** general reference made to specific deterrence and general deterrence

- Offence: weapons dangerous; threaten bodily harm
- Sentence: 5 months probation

---

**ID# 79 - Defense:** "s.24 of the YOA says fail as a last resort and use the least restrictive measures"

- General deterrence - C+, D- J
- Rehabilitation - D+, J
- Specific deterrence - D+, J
- Least restrictive measures - D+

---

**General principles of sentencing - C, D**

**General YOA principles - D+**
<table>
<thead>
<tr>
<th>ID#91 - Judge: general reference by judge to specific and general deterrence</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>offence:</strong> threatening</td>
</tr>
<tr>
<td><strong>sentence:</strong> conditional discharge - 12 months probation, restitution $150</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>ID#92 - Judge: general reference to general &amp; specific deterrence</th>
</tr>
</thead>
</table>
| **Judge:** "Young men don't have the right to carry weapons and wound, harm and possibly kill others. He doesn't express any remorse, takes no responsibility by blaming it on his friend for giving him the knife."
| **offence:** carry concealed weapon |
| **sentence:** 12 months probation, 40 hours CSO with conditions - one of which is to write a 4 page essay which judge itemizes: judge - 4 page essay must be delivered in 15 days:
1 page - respect for rights & integrity of others - must realize that no one is to be threatened by others as you are not to be 2nd page - dangerousness of gang activity and why it should not be tolerated - have to take this seriously - will not accept it otherwise - if you were attacked by a gang you wouldn't like it 3rd page - racial tolerance 4th page - my responsibility as a citizen to my fellow citizens "The purpose of this is not to punish you, it's to get you to think about your actions. You're lucky your lawyer kept you out of jail!" |

<table>
<thead>
<tr>
<th>ID#93 - Crown - &quot;as a matter of law he's responsible - and the fact that other courts didn't attribute restitution (to the co-accused) doesn't mean that the victim should be out $2000 - it's no consolation to the victim that there should be no restitution - different principles of sentencing should apply to restitution - he has benefited equally as others&quot;</th>
</tr>
</thead>
</table>
| **Defense:** "he did not benefit equally as others - consistency of sentencing should be maintained - we have a letter of apology to the victim - my client shouldn't be treated differently when those who got the proceeds didn't have to make restitution"
| **Crown:** "if he's a party to the offence - he's a party to the loss - if he's a party to the loss then he should have to pay. It would be contrary to public policy for the Crown to abide by some arrangement made by a bunch of criminals. "He shouldn't be able to get away with not having to make restitution just because the others didn't have to"
| **Defense:** this is a young offender court - he's a young offender not a criminal - that's what the YOA says - we have to show the young offender that their case is dealt with on principles and there must be a principle of consistency" |
| **Crown:** "this is a restitution issue - not a sentencing issue."
| **offence:** break & enter; theft under sentence: 12 months probation; 100 hours CSO |
### Appendix E - Detailed information on references to S. 24 of the YOA for cases that were considered for custody

<table>
<thead>
<tr>
<th>ID#</th>
<th>Reference to s.24(1) of the YOA?</th>
<th>PDR in case?</th>
<th>Information from PDR presented in court hearing?</th>
<th>Refer to s.24(1) in PDR?</th>
<th>Comments:</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>cases that received custody:</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>no</td>
<td>yes</td>
<td>no</td>
<td>no</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>no</td>
<td>waived</td>
<td>no</td>
<td>no</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>no</td>
<td>waived</td>
<td>--</td>
<td>--</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>no</td>
<td>no</td>
<td>--</td>
<td>--</td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>no</td>
<td>yes</td>
<td>yes</td>
<td>no</td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>no</td>
<td>yes</td>
<td>yes</td>
<td>no</td>
<td>defense - &quot;give him 4 months O/C - that's the going rate on a plea&quot;</td>
</tr>
<tr>
<td>19</td>
<td>?</td>
<td>yes</td>
<td>yes</td>
<td>no</td>
<td>defense - &quot;consider principles of the YOA - extend a period of probation if you're concerned with specific deterrence&quot;</td>
</tr>
<tr>
<td>20</td>
<td>?</td>
<td>no</td>
<td>--</td>
<td>--</td>
<td>defense - &quot;consider absolute discharge because there is no criminal record&quot;</td>
</tr>
<tr>
<td>22</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>no</td>
<td>defense - &quot;any sentencing principles can be met by the maximum hours of community service&quot;</td>
</tr>
<tr>
<td>32</td>
<td>no</td>
<td>yes</td>
<td>no</td>
<td>do not have PDR</td>
<td></td>
</tr>
<tr>
<td>33</td>
<td>no</td>
<td>yes</td>
<td>yes</td>
<td>no</td>
<td></td>
</tr>
<tr>
<td>35</td>
<td>no</td>
<td>yes</td>
<td>no</td>
<td>do not have PDR</td>
<td></td>
</tr>
<tr>
<td>36</td>
<td>no</td>
<td>yes</td>
<td>no</td>
<td>no</td>
<td></td>
</tr>
<tr>
<td>39</td>
<td>?</td>
<td>yes</td>
<td>no</td>
<td>no</td>
<td>defense - &quot;has supportive family and resources in the community - letters from different agencies...I want s/c to send a message of deterrence and o/c to rehabilitate&quot;</td>
</tr>
<tr>
<td>40</td>
<td>no</td>
<td>yes</td>
<td>yes</td>
<td>no</td>
<td></td>
</tr>
<tr>
<td>41</td>
<td>no</td>
<td>yes</td>
<td>no</td>
<td>no</td>
<td></td>
</tr>
<tr>
<td>46</td>
<td>no</td>
<td>no</td>
<td>--</td>
<td>--</td>
<td></td>
</tr>
<tr>
<td>47</td>
<td>no</td>
<td>no</td>
<td>--</td>
<td>--</td>
<td></td>
</tr>
<tr>
<td>54</td>
<td>no</td>
<td>no</td>
<td>--</td>
<td>--</td>
<td></td>
</tr>
<tr>
<td>64</td>
<td>no</td>
<td>waived</td>
<td>--</td>
<td>--</td>
<td></td>
</tr>
<tr>
<td>78</td>
<td>no</td>
<td>waived</td>
<td>--</td>
<td>--</td>
<td></td>
</tr>
<tr>
<td>79</td>
<td>yes-explicit</td>
<td>yes</td>
<td>yes</td>
<td>do not have PDR</td>
<td>defense - &quot;s.24 of the YOA says jail should be used as a last resort...should use least restrictive measures&quot;</td>
</tr>
<tr>
<td>83</td>
<td>no</td>
<td>yes</td>
<td>no</td>
<td>no</td>
<td></td>
</tr>
<tr>
<td></td>
<td>cases that did not receive custody with a PDR:</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>51</td>
<td>no</td>
<td>yes</td>
<td>yes</td>
<td>no</td>
<td>recommend was for custody</td>
</tr>
<tr>
<td>77</td>
<td>?</td>
<td>yes</td>
<td>yes</td>
<td>no</td>
<td>recommend was for community disposition judge - &quot;the most compelling factor in this case is your 3 guilty pleas...by recognizing that you've committed a criminal act, you've taken responsibility...I also consider as a mitigating factor, your supportive family; this is a very significant factor as many youths don't have this support&quot;</td>
</tr>
<tr>
<td>ID</td>
<td>Sex</td>
<td>Age</td>
<td>Status</td>
<td>Present Charge(s)</td>
<td>Previous Convictions</td>
</tr>
<tr>
<td>----</td>
<td>-----</td>
<td>-----</td>
<td>--------</td>
<td>-------------------</td>
<td>----------------------</td>
</tr>
<tr>
<td>016</td>
<td>M</td>
<td>17</td>
<td>(18+)</td>
<td>Possessio...</td>
<td>Cocaine</td>
</tr>
</tbody>
</table>

(1) Probation recommendation - Judge does not agree (2) Y positive reference - 2, Y negative reference - 1, Y youth/Adult reference - 2, A adult references - 1
<table>
<thead>
<tr>
<th>ID</th>
<th>Age</th>
<th>Sources of Info</th>
<th>Categories &amp; Comments</th>
<th>Overall &amp; Recommendation</th>
<th>Sentence</th>
<th>in-court info?</th>
</tr>
</thead>
<tbody>
<tr>
<td>22</td>
<td>14</td>
<td>- theft under - trafficking narcotics (Glasgow with a weapon; 02 in probation &amp; 80 hrs. CSO)</td>
<td>- youth - p.i. - parents - p.i. - program coordinator of Central To Yth Services - principal of high school - probation records</td>
<td>family background: - family constellation &amp; occupations - family history: mother left school at 17 and worked 1/2 at current employer for last 28 years - married &amp; had child with 2nd husband who died - married current partner - yth’s father the young offender: - personality, character, development - mother: no difficulties - with pregnancy and delivery of youth - and youth met all “milestones” at an appropriate rate - raised childhood for youth - except might be on &amp; difficulty with babysitter - defiance: used to be very minor - not picking up after herself - later: more serious - talk back to parents, did not complete school assignments; spent too much time on phone &amp; with friends - whom are older - youth presents as older than thair &amp; engages in activities more appropriate to older youth - ie. all night parties but mother says: she was like that too and so was youth’s older sister - hooked and acted older than age - Mr. X - says youth manipulates her mother - he has history of alcohol abuse and has had treatment - parents have lost trust in daughter: drug/alcohol use, poor school attendance, money missing - home - slow in mornings - late for school - also youth has lost motivation to go to see CAnes - initially eager to attend youth: descent relationship with mom &amp; dad - misses her sister - who she confided in - sister recently married</td>
<td>categories/comments cont’d- - hooked into her peer grp and culture and lacks reinforcement from home - Response to output: weak but has stopped since 1st arrest - since does’t want conflict w/ law - not “moral” reasons - admits smoking my - didn’t realize consequences of offending when 1st caught - now doesn’t do this - wants no trouble w/ law - smokes 5 joints daily - doesn’t believe she has a drug problem - experimentally uses crystal meth, ecstasy, acid and codeine - goes to all night “rave” parties - also panhandles or shames for extra money to attend raves; parents don’t like her going - but do not prevent her</td>
<td>23 days O/C 12 months probation conditions: - no possession of drugs - attend counselling</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

"I realize that you’re a very young girl - but what you did was a very bad thing."

"you’re lacking some serious moral sense."

"going to rems is extremely services."

"today you stop your drug problem"
<table>
<thead>
<tr>
<th>ID</th>
<th>Sex</th>
<th>Age (perceptual age)</th>
<th>(1) Present charges (2) Previous convictions</th>
<th>Sources of Info</th>
<th>Categories &amp; Comments</th>
<th>Overall &amp; Recommendation</th>
<th>Sentence</th>
<th>In-court info?</th>
</tr>
</thead>
</table>
| 8  | m   | 17 (18)             | (1) Robbery - FTC, tire theft under 5000 | - Youth in custody - in residential care - -unt guidance - counselor - correctional officer Metro West - probation officer (ph 1) - probation records - arresting officer - victims | victim comments: - members of community are fearful & have curtailed normal activities. 
- history since Feb 17, 1995 
- family history: resides with mother 
- youth was as a result of a rape when mother was 18 yrs old at a drug party she was attending; therefore no contact w/ biological father 
- home & environment: mother was evicted from home of 3 yrs. Due to illegal act by friend of youth - now mother is housed in motel w/ her 3 children 
- mother is on Mother's allowance and is waiting for O.C.A. Housing 
- stepfather - is in Jamaica - he has contact by phone. was deported to Jamaica for drug trafficking, but youth was not exposed to this. 
- parental influence & degree of control: mother is supportive, youth gets along with stepfather - co-operative at home for most part 
- mother: "not a bad kid" 
- has a curfew - calls if he is out late 
- mostly obeys house rules the young person education (lais subjects & marks) 
- before-distracted other students - poor attendance 
- offender states he has stopped going to school sometimes to avoid certain students/teachers. has difficulty w/ authority figures 
- in alternative learning program - suspended twice - although now described as bright and does well in math 
- teachers: "he is wasting his talent" worked p/t as stock clerk and cleaner. | categories/comments cont'd. - Leisure activities: 
- before: no motivation - watches TV 
- now: recreationally - offender enjoys playing keyboard, watches sports, read, play cards, ride bike - states he drinks alcohol on occasion - admits snorting mj but not cocaine - has no financial debts - has asthma. 
- overall: has not responded well to community supervision in the past - has not complied with probation conditions - offender tells writer that in custody he gets along quite well and follows the program. 
- re: present offence: was cooperative w/ police - regrets involvement w/ law and accepts responsibility for his actions 
- admits that he has some older friends who may have influenced him in the sense that they have material goods that he does not have. 
- recommendation: probation supervision does not meet the needs of the offender at the present time. |
|    |     |                     | (5 yrs. addition to 35 days PTO)          | FTC probation, 75 hrs CSO | 30 days O/C (in addition to 35 days PTO) 
- probation 
- not in possession of non-med drugs 
- no keeping company w/ victims 
- reside with... report to probation officer as new boundary. 
- no pager 
- counselling compensation to victim: $100 - within 12 months. | --- | 
|    |     |                     |                                           |                            | --- | |
| ID | Sex | Age (Prev. Charge) | Present Charge(s) | Sources of Info | Categories & Comments | Overall & Recommendation | Sentence | Inc. Issue?
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>51</td>
<td>m</td>
<td>16 (17)</td>
<td>possession, theft</td>
<td>youth, mother</td>
<td>victim comments</td>
<td>categories/comments/conf</td>
<td>12 months probation &amp; 25 hours community service (joint submission)</td>
<td>mother trying to get youth a 'father-fig'</td>
</tr>
<tr>
<td>#</td>
<td>Sex</td>
<td>Age (Per 16)</td>
<td>Previous charge(s)</td>
<td>Sources of Info</td>
<td>Categorics &amp; Comments</td>
<td>Overall &amp; Recommendation</td>
<td>Sentence</td>
<td>In-court info?</td>
</tr>
<tr>
<td>-----</td>
<td>-----</td>
<td>-------------</td>
<td>-------------------</td>
<td>----------------</td>
<td>----------------------</td>
<td>--------------------------</td>
<td>----------</td>
<td>---------------</td>
</tr>
<tr>
<td>7</td>
<td>m</td>
<td>15 (16)</td>
<td>theft under: FTC with bail; FTA in court</td>
<td>-grade 8 report cards -extra police synopsis of offence -court make-up report -group home report -interview with youth w/ mother</td>
<td>response to present offenses: youth cannot explain his criminal activity - he had no criminal intentions - FTA with bail; FTA in court</td>
<td>categories/comments cont'd - education/employment: -grade 9: has a &quot;C&quot; average; late for class -youth easily bored at class -marks have ranged from 5's - E's previous/current agency info: -has spent the last year in and out of detention -group home report - positive: youth wants out of detention because he misses family &amp; friends -wants to be w/ mom to do chores -dislikes associating with criminals &amp; wants be w/ good friends interview w/ young person: says he feels remorse for his actions -wants to finish school &amp; get a p/t job wants to work w/ computers -spare time: friends, watching TV, babysitting younger siblings, repairing TV's, VCR's and cars &amp; family outings overall -first time offender; 15 years old -offences of theft under: FTC and FTA</td>
<td>-40 hrs of community service work -18 months probation w/ conditions set out in FDR -35 days pre-trial detention to each count (time served)</td>
<td>-Crown: one of the outstanding charges is robbery - set for trial -judge: &quot;most compelling factor in this case is your 3 guilty pleas&quot; -&quot;by recognizing that you've committed a criminal act -you've taken responsibility -also a mitigating factor is your supportive family - this is a significant advantage - many yrs don't have this support&quot; -&quot;you're a 1st time offender -your acts were crimes of opportunity and done of poor judgement - the other two charges were crimes of disrespect of court orders&quot; principles of sentencing mentioned: -p.3's need to be held responsible -can protect society via rehab.</td>
</tr>
<tr>
<td>ID #</td>
<td>Sex</td>
<td>Age (For present charges)</td>
<td>(1) Present charges (2) Previous convictions (dispositions)</td>
<td>Sources of Info</td>
<td>Categories &amp; Comments</td>
<td>Overall &amp; Recommendation</td>
<td>Sentence</td>
<td>In-court info?</td>
</tr>
<tr>
<td>------</td>
<td>-----</td>
<td>------------------------</td>
<td>-------------------------------------------------</td>
<td>----------------</td>
<td>----------------------</td>
<td>------------------------</td>
<td>---------</td>
<td>--------------</td>
</tr>
</tbody>
</table>
| 36 m | 15 (18) | -break, enter & commit theft  
(2) robbery  
(time served: 1.5 months; open custody: 4.5 months; probation -18 months) | -view of probation file  
-meeting w/ mother  
-telephone interview w/ school  
-meeting w/ youth | response to present & previous dispospositions: (nothing written...)  
victim interview: victim is his maternal aunt  
youth had been living w/ her prior to the offence  
had been asked to leave since wasn’t following rules of the house  
-resident home with an elderly - liquor was consumed -CD player & tapes were taken - picture was damaged -most of the loss was recovered by insurance -victim is estranged from youth -he has never apologized -he feels he needs help -but wishes no further involvement  
interviews w/ parents and/or extended family  
-uncle -employed at Toronto hotel  
father - in Jamaica  
half-brother - younger in school  
martial history: mother & her 2 children came to Canada in 1994 -youth’s father never part of the family -but father of half-brother lived with them on & off in Jamaica  
-home & environment: 2 bedroom apartment in regent Park area  
parental influence & degree of control:  
-no problems w/ youth until last year -not listening -would stay out late and sometimes all night  
-has a temper -has damaged some of her property at home during arguments  
-has been in detention & custody due to his offenses  
-mother -not adverse to having him at home -but feels unable to deal with disruptive behavior  
education/employment:  
current school vice-principal: attendance and behavior unacceptable -did better at his other school -in detention-school is a strength | categories/comment/conf'd agencies: parallel assessment being done by Family Court Clinic  
-seen recently @ hospital for sick children since mom thought he was doing drugs -tests turned out negative  
-has epilepsy  
interview with young person  
-sensitive youth whose current situation may be partly explained by medical factors  
-mother: worried about peer influences -drinking, health concerns & lack of positive male model in life  
epileptic - on medication  
-birth & early development were normal -but youth was in a car accident at 9 and suffered serious head injuries young offender: says he really didn’t break in since he was in possession of the key -would rather live w/ his aunt  
doesn’t know how court should deal with him  
-has some interest in sports -but no elaboration on leisure activities  
-has done well in custodial detention  
-has made suicidal gestures w/ medications  
overall:  
-12 year old second offender -whose situation is being complicated by health and familial issues  
recommendation:  
since an 18 month probation order w/ strict conditions made in June—perhaps an additional short custodial disposition followed by an expectation that youth will perform community service would cover basis -and make him feel more accountable to the community | 30 days open custody  
-18 months probation w/ conditions: report as required -attend an assessment by Family Court Clinic  
Crown: not opposed to a/c in light of PDR  
broke trust of aunt  
Defense: suggesting O/C  
judge: agree with Crown |
<table>
<thead>
<tr>
<th>ID#</th>
<th>Sex</th>
<th>Age (Percyman</th>
<th>(1) Present</th>
<th>Sources of</th>
<th>Categories &amp; Comments</th>
<th>Overall &amp;</th>
<th>Sentence</th>
<th>In-court note?</th>
</tr>
</thead>
<tbody>
<tr>
<td>19</td>
<td>m</td>
<td>16(15)</td>
<td>-possession over X3</td>
<td>-step-mother wants youth to live w/ her - but yth says he wants to live w/ family friend because of problems he has w/ his older brother education/employment passed every grade up until 9 w/ average marks - will have to repeat 9 - frequently skipped, not employed - step-mother; he has a learning problem - reading prevent agency involvement; none except probation interview w/ young person: -disservice &amp; liable young ass - when interviewed previously in Metro detention centre was quite cheerful and optimistic - &quot;it wasn't bad&quot; in the institution - yth enjoys drawing, making crafts, building models - has good health &amp; doesn’t use drugs or alcohol - smokes cigarettes - has many friends - some of who have been in trouble w/ the law overall - youth was on probation when he committed these offences - he was 'angry w/ everything' but is not angry now - has never been a problem at Metro West detention centre - when asked what court will do - youth says he doesn’t know - but expects custody - understands the loss he has caused to victim but has no means to make compensation - has been in secure detention for at least 4 months recommendation: - youth has matured over the past few months - would benefit from more probation supervision - if court order custody - recommended that it be short term of open custody which would facilitate his returning to school in the fall</td>
<td>-56 days O/C -12 months probation w/ cond: - report to probation officer - attend program related to driving offences - no learning to drive - curfew 11pm-6am (7 days/wk) - after probation - not to obtain a driver’s license until course completed - reside at approved address - attend school regularly or seek/maintain employment - don’t possess tools for house-breaking</td>
<td>D: family breakdown w/ alcoholic young ass - marital stress</td>
<td>18 months s/c - done &quot;hard time&quot; - 18 months is harsh &amp; not in keeping w/ rehab - consider general &amp; specific deterrents - judge what about safety of public? These are most serious/dangerous offences y/p criminal - D: suggest short period of O/C &amp; extend potential of concern with specific deterrents - he was influenced by an adult who is in jail in Mexico Crown: rehab is important but social demeure - &quot;I'm giving you a chance&quot; &quot;if involved cars again will spend time in jail&quot; - &quot;pos ript form prob officer - says you're matured in custody&quot; &quot;considering served long time in s/c &amp; never served tune before&quot; - &quot;taken chance on you because you plead guilty to 6 charges&quot;</td>
<td></td>
</tr>
<tr>
<td>ID</td>
<td>Sex</td>
<td>Age (Per age)</td>
<td>(1) Present conviction(s) (dispositions)</td>
<td>Sources of Info</td>
<td>Categories &amp; Comments</td>
<td>Overall &amp; Recommendation</td>
<td>Sentences</td>
<td>Ac-count info</td>
</tr>
<tr>
<td>----</td>
<td>-----</td>
<td>---------------</td>
<td>----------------------------------------</td>
<td>----------------</td>
<td>----------------------</td>
<td>------------------------</td>
<td>-----------</td>
<td>---------------</td>
</tr>
<tr>
<td>3</td>
<td>m</td>
<td>17 (16)</td>
<td>-robbery X4 -fail to comply -dangerous driving -weapons -dangerous weapon -setting fire to other substance -masquerade -property (15 months probation) -assault w/ a weapon (30 days O/C) +12 months probation + 75 hours (CSD) -theft over -fail to comply w/ probation (10 days S/C) +12 months probation) -most recent youth -youth's mother -youth's father -youth's friend -family -friend -family -friend -family -counselor (A.R.C) -counselor (Central Toronto Youth Services) -youth's therapist -youth's child service -phase I probation officer -COMSOC -COMSOC records -Corrections records -Police synopsis -victim's statements</td>
<td>particulars of offence -youth rejects some details in police synopsis -but prepared to assume responsibility -wants court to know he knows right from wrong; feels remorse -states another co-accused (he won’t identity) instrumental in the robbery (who wasn’t caught) &amp; one of the co-accused’s caught -was wrongly attributed with the actions of this other person -says robberies weren’t planned &amp; then changed thus -says they were -understands how this affected lives of his family -but not victims victim's comments -all adolescents -more fearful than previously -didn’t want their names in report for fear of retaliation -3 of 4 victims hoped offender would be enrolled as part of his sentence in a rehabilitative program to prevent further victimization family &amp; personal history -parents had an 'awful marriage' -mother -met husband B at 21, trying to escape from her 'domineering' father -father was older and stable -had a job as salesman at furniture mgz co. -she worked as secretary -were able to establish an affluent lifestyle in Forest Hill area of Toronto -marriage broke down -father -verbally abusive &amp; demanding -another -primary parent -but father continued statements of demotion to her and the children -age 3 offender started acting out due to conflict in the home -father ridiculed son (re: something he has talent in -athletics) -age 10 -father left and took up with his current girlfriend -still paid support and tried to control -at same time as father left -family therapist left practice -theme of abandonment in youth's life -another had ref'ship with another man w/ kids who also took off -youth found 'pseudo-family' thru other kids who were macho, defiant and anti-social -parents sent him to private schools in U.S. -where he introduced to gang culture of Harlem &amp; US cities -enrolled back to Toronto with a new persona -hip-hop music; street liaisons; speaks w/ accent -black urban street gang -he is a follower -not a leader -mother: correctional efforts -has daughter -going to university education: did well in sports but his behavior was a concern @ Forest Hill public school -sent off to Private Academy in Georgia -academic setting for kids w/ behavioral problems -then a 6 week wilderness program in Idaho -went to the school in Maine boarding school w/ therapy -returned &amp; started Alt school-behaviour &amp; academic record improved-then went down again when he enter main stream school -is 2 years behind peers in school-wants to go to university</td>
<td>categories &amp; comments -conf'd -employment -worked p/t odd jobs &amp; for his father's furniture company -has had money he needs -mother unsure why he robbed when he has never lacked for things -therapeutic perspective -has now been seeing Dr. X -who says youth is bright, free of mental illness 2 'outrageously unopportunistic' in attitude and presentation. -also seeing a therapist from Jewish child &amp; family services Correctional Supervision: COMSOC report -made acceptable progress: 'moderate' ranking on the risk assessment scale Corrections -reporting has been acceptable -but he gives priority to his social engagements and as late -has only completed a small portion of his CSO at a daycare (which was positive when he went) -but relatively compliant w/ other conditions Previous/current community options/alternatives -not filled out Character, behaviour &amp; attitude -somewhat abusive &amp; anti-authoritarian -but likeable &amp; good potential for future pro-social endeavours -mother: reasons for disruptive behaviour is due to early family life &amp; abandonment by father, school authorities and therapists. -she remains supportive -&amp; has seen some improvement father: describes son as &quot;the weakest young man I have ever met&quot; -believes his behaviour is a result of psychiatric difficulties -family friends: he's bright; a good person; but is striving to establish meaningful ref'ship &amp; a menacing father's behaviour is &amp; scaring by father's abandonment future plans -university assessment -17 years old; has pleaded guilty; currently in Metro West detention centre -has problems of impulsivity that impede him from making clear evaluations of his actions &amp; consequences -parents have spent no expense in the past for therapy, private schools etc. but youth has not made these changes -probation officer: &quot;the many ways the offender presents the world with a false self, a caricature of gang life, all poverty &amp; facade. That is suggestive of poor self esteem and underlying vulnerability&quot;. Recommendation -therapeutic efforts -but if custody (open or secure) is seen as necessary -then Central Toronto Yth's should be involved with offender -community supervision does not appear to be successful in deterring youth's behaviour -if probation is appropriate for sentence -treatment, education and a restriction on associates should be included as conditions 50 days S/C -6 months O/C 18 months probation on w/ conditons: -not in possession of weapons -reside where prob officer specifies -report once a week -obtain counseling from CTYS</td>
<td>-most of the offences occurred in one day judge: &quot;not a serious record&quot; -role of offender -less than others -he didn't brandish the knife Crown: none of the victims is in court watching to see what happens mother; (no judge) -our whole family wanted to be here today incl sister -S/C for 'rehabilitation' -custody is toxic for youth -he is unimpressionable Crown: wants one year in total -leaves S/C/O/C to judge -youth: I'm sorry for what I did -I apologize to victim's mom -won't do it again mother: I'm not going to give up -even father is now committed-needs to be integrated into community... (very articulated) Judge: -has a need to act 'macho' &amp; identify w/ criminal groups -imitating father -prior rec: not that lengthy given family prob -has a lot going for him -I want long term S/C -but this is the last chance for rehabilitation -victims say they are afraid to go out -in one year you'll be out of the youth systeem and on your own</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ID</td>
<td>Sex</td>
<td>Age (Per centual age)</td>
<td>Present Charge(s)</td>
<td>Previous Conviction(s)</td>
<td>Sources of Info</td>
<td>Categories &amp; Comments</td>
<td>Overall &amp; Recommendation</td>
<td>Sentence</td>
</tr>
<tr>
<td>----</td>
<td>-----</td>
<td>----------------------</td>
<td>-------------------</td>
<td>-----------------------</td>
<td>-------------------</td>
<td>---------------------</td>
<td>------------------------</td>
<td>----------</td>
</tr>
<tr>
<td>40</td>
<td>m</td>
<td>16 (17)</td>
<td>(1) escape lawful custody X2</td>
<td>(2) theft under 8 months probation + 30 hrs CSO</td>
<td>- youth - youth's mother - youth's father</td>
<td>- court's comments (in police synopsis)</td>
<td>- owes $74.95</td>
<td>1 day S/C (nothing he had been in pre-trial detention for 38 days) - and still has 8 months O/C - will you run away again?</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>- theft under (fine $74.95)</td>
<td>- possession of stolen property</td>
<td>- Tiva/12 months probation on each charge</td>
<td>- family background: married since 1981 - 2 kids - 4-6 bedroom house - offender upbringing was normal but due to hyperactivity had difficulties socializing, w/ other children - dangeous elementary school participated in flowers &amp; Cake and Cakes; played baseball for 6 years - mother: never been violent at home, school or community - mother: O/C not restrictive enough - because he is impulsive - will escape - needs a strict program to motivate him - offender is a follower; lacks self-confidence - mother wants him to go to school 1/1 or get employment - father: son would be a good candidate for 'boot camp' - because he is non-violent but needs regimentation - father: has a disregard for authority; an active liar; father feels no influence over son - has not visited him since he was incarcerated. Youth lacks respect - has A.D.D. &amp; was put on ritalin at age 7 - but taken off it when he was found giving them to friends - no other evidence of drug/alcohol abuse - youth - expresses remorse education - probe began in kindergarten - concentration was poor - low achievement - disruptive in class in high school - refused to attend classes</td>
<td>- CAS - foster care &amp; request of parents employment - operates a newspaper route for the last 2 years -74 customers - shop boy at department store earning $6.62/hr - 3 months job due to absenteeism previous present community options - phase 1 probation officer: youth unable to function normally outside of a non-structured setting - poor candidate for community disp. - if O/C - he will likely 'bott' again - advises a Youth camp setting open custody behaviour - positive until incident occurred institutional/behaviour - satisfactory character, behaviour &amp; attitude - youth: were 3 reasons he left O/C: (1) he wasn't allowed to participate in programs his first week there (2) the house was dirty (3) he had difficulty making friends future plans - finish school; move in w/ parents; attend counselling &amp; get a p/t job overall - age 16, repeat offender; shows remorse - admits lying, stealing &amp; being disrespectful towards his parents are areas he needs to work on - family is caring &amp; has tried to get counselling &amp; intervention for him recommendation - if a period of community supervision is seen as appropriate - probation conditions should address regular school attendance; non-association w/ those w/ a criminal record; and counselling as directed</td>
<td></td>
</tr>
<tr>
<td>Index</td>
<td>Sex</td>
<td>Age (FARC situations, 1 day)</td>
<td>Present charge(s)</td>
<td>Previous conviction(s) (dispositions)</td>
<td>Sources of Info</td>
<td>Categories &amp; Comments</td>
<td>Overall &amp; Recommendation</td>
<td>Sentence</td>
</tr>
<tr>
<td>-------</td>
<td>-----</td>
<td>-----------------------------</td>
<td>------------------</td>
<td>-------------------------------------</td>
<td>---------------</td>
<td>----------------------</td>
<td>------------------------</td>
<td>----------</td>
</tr>
<tr>
<td>1</td>
<td>m</td>
<td>16</td>
<td>trafficking</td>
<td>cocaine</td>
<td>-interview w/ youth</td>
<td>victim's background n/a</td>
<td>-16 year old first offender</td>
<td>2 months secure custody + 12 months probation w/ conditions: -report every 2 weeks -receive where directed -obey written rules of the house -attend school during term of probation</td>
</tr>
<tr>
<td>Name</td>
<td>Sex</td>
<td>Age (Percep. Age)</td>
<td>Sources of Info</td>
<td>Categories &amp; Comments</td>
<td>Overall &amp; Recommendation</td>
<td>Sentence</td>
<td>In-Court Info?</td>
<td></td>
</tr>
<tr>
<td>------</td>
<td>-----</td>
<td>-------------------</td>
<td>----------------</td>
<td>-----------------------</td>
<td>--------------------------</td>
<td>----------</td>
<td>---------------</td>
<td></td>
</tr>
<tr>
<td></td>
<td>m</td>
<td>16 (17)</td>
<td>(1) Present charge(s): possession over (stolen vehicle)</td>
<td>youth - personal interview w/ vice-principal of high school, - case summary notes of youth at CAS (1981-1994) - case file - probation &amp; community services 1994-1997 - case conference notes - St. John's school</td>
<td>categories/comments conf'd the young offender - personality &amp; development - bright &amp; personable. Describes ref. ship w/ mother in positive terms - protective of his mother - reasons fathers abuse and neglect - and upset since father is making a greater effort to parent his new son w/ current partner - he is interested in pursuing his native heritage &amp; spirituality - has done well on probation - he is a very &quot;thoughtful &amp; insightful youth w/ a keen awareness of the limitations that his life in relative poverty have brought&quot; - responding well to structure of program @ St. John's school - sister visited - minimal contact from own education; behavioural problems at school - lack of participation by mother in interventions - obstacle for him to get appropriate programming - absence - vice-principal was aware of youth's aggression but found him to be intelligent &amp; likable - s.p. doesn't think youth should return to this highschool - he has a bad reputation there - needs a fresh start</td>
<td>2 weeks O/C consecutive to other sentence</td>
<td>- Crown: was in O/C previously for 3 months - this is his 2nd stolen vehicle - was on bail @ the time - has served 91 days S/C - should do more than this now - Defense: mother wants him to serve O/C up North in a native facility - still will be on probation until 18 years old - Judge: has served 91 days - sentence 2 weeks O/C consecutive</td>
<td></td>
</tr>
<tr>
<td>Category &amp; Comments</td>
<td>Overall &amp; Recommendations</td>
<td>Case Study</td>
<td>Sentence</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>---------------------</td>
<td>---------------------------</td>
<td>------------</td>
<td>----------</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>โปรแร้คSUMMARY/COMMUNITY Case Study</td>
<td>Case Study</td>
<td>Case Study</td>
<td>Sentence</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PCA's and other colleagues have been involved in his life, but another followed him into a new world.</td>
<td>PCA's and other colleagues have been involved in his life, but another followed him into a new world.</td>
<td>PCA's and other colleagues have been involved in his life, but another followed him into a new world.</td>
<td>PCA's and other colleagues have been involved in his life, but another followed him into a new world.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PCA's and other colleagues have been involved in his life, but another followed him into a new world.</td>
<td>PCA's and other colleagues have been involved in his life, but another followed him into a new world.</td>
<td>PCA's and other colleagues have been involved in his life, but another followed him into a new world.</td>
<td>PCA's and other colleagues have been involved in his life, but another followed him into a new world.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PCA's and other colleagues have been involved in his life, but another followed him into a new world.</td>
<td>PCA's and other colleagues have been involved in his life, but another followed him into a new world.</td>
<td>PCA's and other colleagues have been involved in his life, but another followed him into a new world.</td>
<td>PCA's and other colleagues have been involved in his life, but another followed him into a new world.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PCA's and other colleagues have been involved in his life, but another followed him into a new world.</td>
<td>PCA's and other colleagues have been involved in his life, but another followed him into a new world.</td>
<td>PCA's and other colleagues have been involved in his life, but another followed him into a new world.</td>
<td>PCA's and other colleagues have been involved in his life, but another followed him into a new world.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PCA's and other colleagues have been involved in his life, but another followed him into a new world.</td>
<td>PCA's and other colleagues have been involved in his life, but another followed him into a new world.</td>
<td>PCA's and other colleagues have been involved in his life, but another followed him into a new world.</td>
<td>PCA's and other colleagues have been involved in his life, but another followed him into a new world.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PCA's and other colleagues have been involved in his life, but another followed him into a new world.</td>
<td>PCA's and other colleagues have been involved in his life, but another followed him into a new world.</td>
<td>PCA's and other colleagues have been involved in his life, but another followed him into a new world.</td>
<td>PCA's and other colleagues have been involved in his life, but another followed him into a new world.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Chapter Four – Public Perceptions

Introduction

The preceding chapters reveal that neither the ‘youthfulness’ of offenders, nor their chronological age appear to be independently related to the outcome of bail or sentencing hearings. Instead, the two court observation studies showed that legal factors accounted for much of the decision-making occurring in youth court. Moreover, provisions specific to the Young Offenders Act were rarely raised in the court hearings (e.g. the ‘responsible person’ provision in bail hearings, S.3 principles of the YOA). These findings lead one to question how the principles that have been written specifically for young persons who come in conflict with the law, come into play in the youth justice process.

The premise for having a separate set of principles for dealing with young people is rooted in the assumption that there are differences in the nature of offending based upon age. Consequently, the Young Offenders Act and its predecessor the Juvenile Delinquents Act, put forth a different set of standards for youthful offenders. Even the recently proposed replacement for the Young Offenders Act has reaffirmed that there should be a separate justice system for youths. However, growing public attention to the problem of crime and particularly to serious youth crime has resulted in a shift in political focus. Rather than childhood being considered a factor which mitigates criminal responsibility, the seriousness of the offence and the risk the offender poses to the community appear to be taking precedence over ‘childhood status’. This kind of approach was illustrated recently by a youth court judge in Florida. In a case of first degree murder involving a 15

---

year old offender, the judge, in passing down an adult life sentence indicated to the youth
"I do not perceive you to be a child... (y)our monstrous act made you an adult".²

Moreover, research has shown that members of the public do not see differences in
the seriousness of crimes based upon the age of the offender. For example, a study done
quite some time ago (Sellin & Wolfgang 1964) provided a series of brief offence
descriptions to subjects in an experiment which sought to understand how the public rated
the seriousness of various kinds of offences. Respondents were comprised of university
students, police officers, members of the public selected for jury duty and juvenile court
judges. In addition to the offence descriptions which included variations in the context of
offences (such as the presence or absence of bodily injury, whether property was stolen or
damaged, or whether intimidation was present etc.), the authors also randomly assigned
information on the age of the hypothetical offender to respondents. In all cases the
offender was male. In the first condition no other information was given about the
offender, and in the other three conditions, respondents were told he was either 27 years
of age, 17 years of age, or 13 years of age. The results showed that age did not affect
judgments of the seriousness of the crime. As stated by the authors:

The age of the offender does not particularly color a person's judgment about
the seriousness of the offence. A pervasive social agreement about what is serious
and what is not appears to emerge, and this agreement transcends simple qualitative
concordance; it extends to the estimated numerical degree of seriousness of these
offences (Sellin & Wolfgang 1967: 268).

Based upon findings such as these, some researchers have concluded that there need not
be a separate justice system for youths since "...the likelihood of crime varies continuously

² "15-year-old sentenced to life" The Toronto Star August 21ª, 1999
with age, but the meaning of criminal acts does not depend on the age of the offender. Distinctions based upon age are thus arbitrary" (Hirschi & Gottfredson 1993).

The purpose of this chapter is to examine if the public makes distinctions among offenders based upon apparent age and maturity. In light of the court observation findings, it is important to understand how the public responds to the question of age given that court practitioners may feel pressured to reflect public sentiment in making courtroom decisions (Ouimet & Coyle 1991). I will examine this question by investigating two related issues; first, since the ‘youthfulness’ of an offender was not independently related to court decisions (as seen in the previous chapters), I will examine how information on the youthfulness of a young offender may affect public preferences. Specifically, if a young offender is constructed to be seen in more ‘youthful’ or more ‘adult-like’ terms, does this affect public responses regarding the characteristics of that offender, and the kind of sentence which is seen to be most appropriate in the case?

Second, the court observation studies also revealed that principles and goals of sentencing that were specific to young people in conflict with the law were rarely raised in youth courts. Instead, goals such as general and specific deterrence and issues of public protection were more frequently mentioned in sentencing hearings. Consequently, is there reason to believe that the public differentiates goals of sentencing based upon whether the offender is a youth or adult, or are the purposes of sentencing seen as the same regardless of the age of the offender?

---

3 Ouimet & Coyle (1991) found that court practitioners’ perceptions of the public’s fear of crime had an impact on their own decisions in less serious kinds of cases.
**Research Methods**

During the summer of 1997, a public attitudes survey was carried out in Ontario. This project was a result of a cooperative agreement between Operation Springboard and the Centre of Criminology at the University of Toronto. The actual survey itself was carried out by Goldfarb and Associates in Toronto, Ontario.

Using a random digit dialing technique, 1006 households across Ontario were successfully contacted and one adult within each household was interviewed. Interviews were carried out in English only. Approximately half (n=500) of the respondents were asked questions which dealt largely with adult crime and the criminal justice system. The remainder (n=506) were asked questions about youth crime and their views regarding the youth justice system. In many cases equivalent questions were asked of respondents in the two groups which allowed for comparisons in reaction to youthful and adult offenders and the separate justice systems.

**Section 1 - Examining the effects of information about young offenders - more 'youthful' vs. more 'adult-like'.**

Within the youth justice system there is a recognition that age and maturity are factors which need to be evaluated in determining the most appropriate outcome for a case (Section 3(c) YOJA). While court practitioners may be trying to interpret how this section of the legislation should affect court decisions, presumably they may feel compelled to reflect public sentiment in dealing with cases. Thus, there is an interdependency between

---

4 Funding for the project to Operation Springboard was provided by the Trillium Foundation and Bell Canada. The survey itself was carried out at a reduced cost by Goldfarb consultants.

5 Operation Springboard is a non-profit community organization which provides correctional services to both adult and young offenders in Toronto and other areas in Southern Ontario.

6 Details can be found in Doob A.N., J.B. Sprott, V. Marinos and K. Varma (1998). *An Exploration of Ontario residents’ views of crime and the criminal justice system*. Toronto: Centre of Criminology.
the courts and the public (Kaukinen and Colavecchia 1999). For the purposes of this research then, it is important to find out how the public responds to differing constructions of age and maturity among young offenders. An experiment was conducted within this survey to assess how information about the ‘youthful’ or ‘adult-like’ characteristics of a young offender may affect responses of the public to a case. This experiment was only given to those responding to the youth survey (n=506). The question started out by describing the following scenario:

Imagine the following case. A 17 year old male young offender is found guilty of stealing a car and driving it around for a couple of hours before being involved in a minor accident where nobody was hurt.

There were 5 different conditions in this experiment. Each of 5 different randomly assigned groups were given slightly different pieces of information about the young person. The first condition had no extra information. Each of the other descriptions were designed to portray the youth on a continuum from more ‘youthful’ all the way up to more ‘adult-like’ descriptions. My hypothesis was that the public would respond more favourably, and therefore, less punitively to the most ‘youthful’ characterization of the offender. The 5 conditions were as follows and the variations in description have been italicized:

(1) no additional information

(2) He is about 5 feet 10 inches tall, with dark wavy hair, and is seen as being moderately attractive. He was quiet in court and gave no clear explanation for the offence. *Both of his parents were in court with him but they did not say anything at the court hearing.* The probation officer reported to the court that *he attends school regularly; he is doing average work in school,* and generally that people who know him see him as quite an ordinary person.
(3) He is about 5 feet 10 inches tall, with dark wavy hair, and is seen as being moderately attractive. He was quiet in court and gave no clear explanation for the offence. *He came to court alone; his parents were not apparently with him.* The probation officer reported to the court that he attends school regularly; he is doing average work in school, and generally that people who know him see him as quite an ordinary person.

(4) He is about 5 feet 10 inches tall, with dark wavy hair, and is seen as being moderately attractive. He was quiet in court and gave no clear explanation for the offence. *He came to court alone; his parents were not apparently with him.* He is receiving welfare and is living alone. The probation officer reported that generally people who know him see him as quite an ordinary person.

(5) He is about 5 feet 10 inches tall, with dark wavy hair and a moustache, and is seen as being moderately attractive. He was quiet in court and gave no clear explanation for the offence. *He is no longer in school; he lives by himself, and is working full-time.* The probation officer reported to the court that he works regularly; *his employer has no problems with him,* and that generally people who know him see him as quite an ordinary person.

When asked which type of sanction they thought would be most appropriate for the young offender in their scenario. They were given the following choices:

- a period of time in custody
- a community service order where he had to work for a certain number of hours without pay for the owner of the car or a community agency
- a fine

**Results: Sentencing preferences**

The hypothesis that respondent’s ratings would vary based upon descriptions of the offender along the dimension of age turned out not to be the case. Instead, the most interesting aspect of these results did not relate to differences in ratings based upon ‘youthful’ or ‘adult-like’ characterizations. Rather, respondents differed in how they rated this young offender based upon receiving *any* information about him at all (Table 1).

---

7 For the purposes of statistical analysis, the sentencing options of the community service order (CSO) and fine were combined and examined in contrast to a sentence of imprisonment.
Table 1 - Sentencing preferences for case scenario as a function of various descriptions of the youth

<table>
<thead>
<tr>
<th>Description of youth at sentencing hearing:</th>
<th>Type of sentence seen as most appropriate:</th>
<th>Totals</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Prison</td>
<td>CRO/FINE</td>
</tr>
<tr>
<td>no extra information</td>
<td>22 (21.4%)</td>
<td>81 (78.6%)</td>
</tr>
<tr>
<td>parents in court, attends school</td>
<td>12 (12.4%)</td>
<td>85 (87.6%)</td>
</tr>
<tr>
<td>no parents in court, attends school</td>
<td>17 (14.8%)</td>
<td>98 (85.2%)</td>
</tr>
<tr>
<td>no parents in court, receives</td>
<td>13 (14.3%)</td>
<td>78 (85.7%)</td>
</tr>
<tr>
<td>welfare, lives alone</td>
<td></td>
<td></td>
</tr>
<tr>
<td>not in school, lives alone, works</td>
<td>5 (5.2%)</td>
<td>91 (94.8%)</td>
</tr>
<tr>
<td>full-time, moustache</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note: Chi-square = 11.219, df = 4, p = .024

*No extra information* vs. the other categories (pooled): Fisher’s exact test p = .016

The preference for a sentence of imprisonment for this youth was chosen most frequently when no extra information about the youth was given to respondents. As Table 1 shows, one-fifth (21.4%) of respondents chose prison for the 'contextless' young offender. The only other apparent difference in the scenarios also ran counter to my hypothesis. Surprisingly, the youth who was purposefully constructed to be most 'adult-like' (condition #5) was least likely to have prison chosen as the most appropriate sanction. The preference for prison in the other 3 descriptions only varied slightly.

What this suggests is that the public is influenced by more information given to them about an offender rather than the qualitative details of that information (with the possible exception of the most 'adult-like' youth). It seems that descriptive information which characterizes the 'young offender' as a 'young person who offended' allows members of the public to move beyond more punitive responses, towards
sentencing preferences which may be more meaningful to the actual circumstances of the offender. This general finding is consistent with previous research on the effects of contextual information on public attitudes.

For example, Doob and Roberts (1983) conducted a study in which 116 Ontario residents were asked to evaluate sentences handed down in two separate kinds of cases -- a manslaughter case, and a criminal negligence case. The authors purposefully chose cases that appeared to have been given mild sentences by the trial court judge and were upheld by the Courts of Appeal. Respondents were given either a short description of the case, or a longer version in which a more complete account of the facts of the case were given. The results of this study showed that when given a longer version which provided information about the surrounding circumstances of a case, respondents were significantly less likely to rate the sentence in the case as being 'too lenient' than those who received a short description of the case. Covell and Howe’s (1996) study put forward similar results regarding the power of information on punitiveness regarding young offenders. The authors carried out a survey of 247 respondents ranging in age from 15 to 45 years old. The questionnaire examined the respondents’ knowledge of the basic provisions of the Young Offenders Act as well as provided case scenarios about serious crimes –murder or sexual assault –perpetrated by a male or female offender. The questionnaires were structured so that respondents were organized to randomly assigned to receive either basic information on the offence and verdict, or basic information along with a paragraph of background information on the offender. The gender and offence type were also randomly assigned. The authors found that across all four conditions, the level of information was the most consistent and greatest predictor or attitudes. Respondents who were given
extra background information about the young offender showed significantly less punitive attitudes compared with those who were given only the basic information on the offence and sentence (Covell and Howe 1996).

Another study conducted by Roberts and Doob (1990) found that subjects’ ratings of court sentences were also affected by the type of information they received about a sentencing case. The authors found that when subjects were given a summarized version of a sentencing hearing as opposed to a media description of the same sentencing case, a significantly smaller proportion of subjects rated the sentence as being ‘too lenient’. In addition, those subjects who were given the summary of court transcripts held less negative views of the offence and the offender than did those who read the media version of the case (Roberts & Doob 1990).

Stalans (1993) also found that subjects’ responses were affected by the information they received. Stalans’ study revealed that providing a realistic stereotype of an offender as opposed to offender stereotypes which are represented through the media, resulted in lower demands for harsh punishments. In addition, this study showed that unrealistic stereotypes about offenders could be reduced by providing contextually distinct information about crime stories involving minor harm (Stalans 1993).

Finally, Lane’s (1997) study on the effects of a correctional course on levels of punitiveness for undergraduate students, reveals that more information about criminal justice appears to be associated with less punitive sentencing preferences. Lane’s study, which assessed responses on a number of hypothetical case scenarios, found that by the end of the course on corrections, the preference for prison was reduced in scenarios
involving non violent offenders but there was less of a change in levels of punitiveness for scenarios involving violent offenders (Lane 1997).

In the context of the present study, an obvious question that arises from the finding that more information about a youth reduced respondents' preference for prison, is the possibility that the effects of more information may not be as important for members of the public who hold more punitive attitudes. Extra information about a youth’s circumstances may have little influence on respondents who believe that sentences in youth court are too lenient. The following table (Table 2) shows the responses to this experiment by only those who responded that youth court sentences are too lenient.

**Table 2 - Sentencing preferences for case scenario as a function of various descriptions of the youth by respondents who perceived youth court sentences to be ‘too lenient’**

<table>
<thead>
<tr>
<th>description of youth at sentencing hearing:</th>
<th>youth court sentences are not severe enough</th>
<th>prison</th>
<th>cso/fine</th>
</tr>
</thead>
<tbody>
<tr>
<td>no extra information</td>
<td></td>
<td>22 (25.0%)</td>
<td>66 (75.0%)</td>
</tr>
<tr>
<td>parents in court, attends school</td>
<td></td>
<td>11 (14.3%)</td>
<td>66 (85.7%)</td>
</tr>
<tr>
<td>no parents in court, attends school</td>
<td></td>
<td>15 (16.9%)</td>
<td>74 (83.1%)</td>
</tr>
<tr>
<td>no parents in court, receives welfare, lives alone</td>
<td></td>
<td>11 (15.3%)</td>
<td>61 (84.7%)</td>
</tr>
<tr>
<td>not in school, lives alone, works full-time, moustache</td>
<td></td>
<td>4 (5.6%)</td>
<td>68 (94.4%)</td>
</tr>
</tbody>
</table>

**Significance:** Chi-square=11.483, df=4, p=.022

NOTE: information is not shown for the group of respondents who thought ‘youth court sentences are too severe/about right’ since 63/65 of these respondents chose a CSO/fine as the preferred sentence.

Once again, the preference for prison decreased when any information was given about the youth, even for those respondents who believed that youth court sentences should be more
severe. This is not surprising since most respondents in this sample believed that youth court sentences are not severe enough and so the sample size in this table is only slightly smaller than the previous one (Table 1). The most ‘adult-like’ yielded the fewest responses for prison. The preference for prison for this youth was substantially lower than in the other three conditions with varying degrees of descriptive information.

**Ratings of the characteristics of this youth:**

In order to understand if the “youthfulness” of an offender affected public responses in other ways, another set of questions asked respondents to rate this young offender on a number of different dimensions relating to his character:

- **dangerousness** 1=not at all dangerous; 10=very dangerous
- **honesty** 1=dishonest; 10=honest
- **maturity** 1=he is young and immature; 10=his is mature
- **employability** 1=not a good candidate for employment; 10=a good candidate for employment
- **crime was intentional** 1=was not thinking & made a mistake; 10=knew exactly what he was doing
- **likelihood of reoffending** 1=very unlikely to reoffend in future; 10=very likely to reoffend in future

Each dimension was described in a way where a ‘1’ indicated that the youth was low in the characteristic and a ‘10’ meant that he was high in that particular characteristic.

Table 3 shows the mean ratings of surveyed respondents based upon the 5 randomly assigned conditions.
Table 3 - How the young person was rated as a function of the information that was given about him

<table>
<thead>
<tr>
<th>Dimension:</th>
<th>Experimental Condition: Description of youth given to respondent</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th>sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>no extra information</td>
<td>parents in court, attends school</td>
<td>no parents in court, attends school</td>
<td>no parents in court, receives welfare, lives alone</td>
<td>not in school, lives alone, works ft, moustache</td>
<td></td>
</tr>
<tr>
<td>dangerousness</td>
<td>5.58</td>
<td>4.01</td>
<td>4.07</td>
<td>4.12</td>
<td>3.65</td>
<td>$F=11.167$ $p&lt;.001$</td>
</tr>
<tr>
<td>honesty</td>
<td>4.06</td>
<td>4.04</td>
<td>4.95</td>
<td>4.93</td>
<td>4.47</td>
<td>$F=3.236$ $p=.012$</td>
</tr>
<tr>
<td>maturity</td>
<td>3.26</td>
<td>3.95</td>
<td>4.55</td>
<td>3.92</td>
<td>3.98</td>
<td>$F=4.844$ $p=.001$</td>
</tr>
<tr>
<td>employability</td>
<td>3.97</td>
<td>4.87</td>
<td>5.56</td>
<td>4.75</td>
<td>5.83</td>
<td>$F=11.702$ $p&lt;.001$</td>
</tr>
<tr>
<td>crime was intentional</td>
<td>5.80</td>
<td>5.68</td>
<td>5.20</td>
<td>5.48</td>
<td>5.08</td>
<td>n.s.</td>
</tr>
<tr>
<td>likelihood of reoffending</td>
<td>6.16</td>
<td>4.65</td>
<td>4.48</td>
<td>5.18</td>
<td>4.71</td>
<td>$F=8.289$ $p&lt;.001$</td>
</tr>
</tbody>
</table>

NOTE: Expressed as a mean score out of 10 for each dimension

There were significant differences in ratings on all of the dimensions except for whether the crime was intentional or not. The ‘face-less’ young offender was rated as more dangerous, less mature, less employable, and more likely to re-offend than the youth described in the other 4 conditions. The ‘face-less’ youth was also rated low on the dimension of honesty - but the offender described in the second condition who had parents in court and was attending school was rated as slightly less honest.

Once again, while there was some variation in ratings within the other descriptive conditions, there appears to be no clear trend in rating the more ‘youthfully’ constructed offender as being different from the more ‘adult-like’ offender. Ratings of this young person on different dimensions did not appear to be contingent upon the actual qualitative

---

8 It is quite interesting to note that respondents were willing to assess these characteristics based upon the small amount of information they received. Obviously they were asked to answer the question whether they may have felt comfortable doing so or not. Nevertheless, it reveals that the public will assess cases with very little information (Roberts & Doob 1990) and this may provide some more insight into the way the public forms ‘offender stereotypes’ from brief sources of information on a case (Stalans 1993).
differences in descriptions. As before, the most consistent effect was the effect of information vs. no information.

The interaction of ratings of the youth and sentencing preferences:

To pursue this question a bit further, an examination was completed in order to understand if there were differences in the ratings of this youth based upon both variations in the information given and by sentencing preferences, (Table 4).

Table 4 - How the young person was rated as a function of the information that was given about him by sentencing preferences

<table>
<thead>
<tr>
<th>Dimension</th>
<th>no extra information</th>
<th>parents in court, attends school</th>
<th>no parents in court, attends school</th>
<th>no parents in court, receives welfare, lives alone</th>
<th>not in school, lives alone, works ft, moustache</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>prison</td>
<td>cso/ fine</td>
<td>prison</td>
<td>cso/ fine</td>
<td>prison</td>
</tr>
<tr>
<td>Dangerousness</td>
<td>7.77</td>
<td>4.99</td>
<td>5.17</td>
<td>3.86</td>
<td>5.65</td>
</tr>
<tr>
<td>Honesty</td>
<td>3.09</td>
<td>4.37</td>
<td>3.50</td>
<td>4.13</td>
<td>2.94</td>
</tr>
<tr>
<td>Maturity</td>
<td>3.64</td>
<td>3.16</td>
<td>5.18</td>
<td>3.81</td>
<td>4.88</td>
</tr>
<tr>
<td>Employability</td>
<td>3.32</td>
<td>4.16</td>
<td>5.33</td>
<td>4.80</td>
<td>4.29</td>
</tr>
<tr>
<td>Crime was intentional</td>
<td>6.50</td>
<td>5.60</td>
<td>7.83</td>
<td>5.44</td>
<td>5.94</td>
</tr>
<tr>
<td>Likelihood of re-offending</td>
<td>7.81</td>
<td>5.71</td>
<td>6.09</td>
<td>4.46</td>
<td>6.19</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Dimension</th>
<th>main effects (varying information)</th>
<th>main effects (preference for prison vs. cso/fine)</th>
<th>2-way interaction effects (varying information &amp; sentencing preferences)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dangerousness</td>
<td>F=9.061, p.&lt;.001</td>
<td>F=38.162, p.&lt;.001</td>
<td>F=2.675, p=.031</td>
</tr>
<tr>
<td>Honesty</td>
<td>F=3.154, p=.014</td>
<td>F=20.810, p.&lt;.001</td>
<td>n.s.</td>
</tr>
<tr>
<td>Maturity</td>
<td>F=4.943, p=.001</td>
<td>n.s.</td>
<td>n.s.</td>
</tr>
<tr>
<td>Crime was intentional</td>
<td>n.s.</td>
<td>F=16.887, p.&lt;.001</td>
<td>n.s.</td>
</tr>
<tr>
<td>Likelihood of re-offending</td>
<td>F=6.809, p.&lt;.001</td>
<td>F=54.128, p.&lt;.001</td>
<td>n.s.</td>
</tr>
</tbody>
</table>

NOTE: 2 way interaction effects occurred for the dimension of 'dangerousness' only. Given that this occurred only for the dimension of 'dangerousness' it is difficult to make any generalizations about the meaning of this effect.
As with the previous analysis, the conditions in which there was an attempt to construct the youth on a continuum of more ‘youthful’ to more ‘adult-like’ (conditions 2 through 5), showed no apparent differences in mean ratings, nor any consistent patterns in ratings between those that preferred prison, and those that chose a CSO or fine for this youth.

However, looking again at the condition with ‘no information’ versus the remaining conditions, shows that there was an effect of information on how respondent’s rated the youth in most cases. Looking at the conditions of ‘dangerousness’, ‘honesty’, ‘employability’ and ‘likelihood of reoffending’ shows that respondents who had no extra information generally appeared to rate this young offender in a more negative way.

Respondents who chose prison as the most appropriate sanction for this offender and did not get extra information also rated him as more dangerous, less employable and more likely to re-offend in the future, than either those respondents who chose prison in the other 4 conditions, or those who chose a CSO or fine and also did not get extra information. What is also quite interesting to note is that those who preferred a CSO or fine for the ‘no information’ offender, also rated him in a more negative way than those who chose a CSO or fine in the other conditions with descriptive information. Thus, those who preferred a CSO or fine and did not get extra information also rated this youth as more dangerous, less employable and more likely to re-offend in the future compared to those who chose a CSO or fine in the other 4 conditions.

Thus, at least on the ratings of dangerousness, employability and likelihood of re-offending, not only is there an effect where those who chose prison were more likely to see the young offender in more negative terms, but also that those who preferred sanctions other than prison were also more likely to see the young offender in a negative way if they
did not have extra information. Extra information had an effect on both groups of respondents—those that preferred prison and even those that preferred a sanction other than imprisonment.

From a policy standpoint, providing more information about young offenders who have broken the law may be important for the public when they are asked to assess the adequacy of youth court sentencing practices. Judges and the public get quite different kinds of information when asked to sentence cases: judges make their decisions from kinds of information as evidenced in the 4 latter conditions of this experiment, while the public are typically asked to make sentencing recommendations from condition 1. Perhaps by being able to ‘put a face on the offender’, they may be in a better position to assess sentencing decisions by having a more complete story. Even certain judges have remarked that when trying to understand what to do in a young offender’s case, they think about their own children of a similar age in attempting to decide on an appropriate sentence.9.

Section II - Does the public differentiate between youths and adults on the purposes of sentencing?

The second part of this analysis also arises out of the results of the previous chapters. As shown in Chapters 2 and 3, legal factors accounted for most of the decision-making occurring in bail and sentencing hearings. Age and ‘youthfulness’ appeared not to have an independent effect on the outcome of cases. In addition, as Section I of this chapter has shown, even the public does not seem to respond differently to characterizations of youth constructed on a continuum of ‘maturity’. The reasons for the

9 From personal conversations with youth court judges
lack of an effect are puzzling, but may be partly explained by the overall shift that seems to be taking place within youth justice systems in most Western industrialized countries. For instance, recent research has noted that there is a persistent erosion occurring to the separate justice system for youth. The dismantling of the ideals of a separate youth justice system, it appears, are occurring within a climate of immense public support (Sprott 1998). The public is said to be less tolerant of a mitigated approach for young people who break the law which has resulted in numerous ‘get tough’ policies for young offenders in recent years.

The results from this survey of residents in Ontario are consistent with a harsh approach to dealing with young offenders. As Table 5 shows, the majority of respondents (63.8%) indicated that there should not be a separate system of justice for youths. A similar proportion (65.0%) of respondents thought that if an adult or a youth committed a similar kind of offence, the youth should receive a sentence that is the same as or harsher than the sentence an adult would receive. Finally, the large majority (85.7%) of respondents felt that youth court sentences were not severe enough. The only area in which the majority of those surveyed appeared to make a distinction between adults and youths was regarding separate prisons for youthful offenders. The majority (86.6%) of respondents thought that youth should be kept in separate facilities.
Table 5 - Percentages showing respondents' views of youth and adult justice

<table>
<thead>
<tr>
<th></th>
<th>oppose/strongly oppose</th>
<th>favour/strongly favour</th>
<th>TOTAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>separate youth justice system?</td>
<td>63.9%(399)</td>
<td>36.1%(339)</td>
<td>100%(938)</td>
</tr>
<tr>
<td>separate facilities for young offenders?</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>same prisons</td>
<td>separate prisons</td>
<td></td>
</tr>
<tr>
<td></td>
<td>10.7%(104)</td>
<td>89.3%(868)</td>
<td>100%(972)</td>
</tr>
<tr>
<td>youth sentences—same, harsher or more</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>same, harsher or more lenient than adult</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>sentences for a property offence?</td>
<td>65.0%(642)</td>
<td>35.0%(346)</td>
<td>100%(988)</td>
</tr>
<tr>
<td></td>
<td>not severe enough</td>
<td>too severe/about right</td>
<td></td>
</tr>
<tr>
<td>sentences in youth court—severe,</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>about right or not severe enough?</td>
<td>85.7%(800)</td>
<td>14.3%(134)</td>
<td>100%(934)</td>
</tr>
</tbody>
</table>

This indicates that on the surface, the majority of people in Ontario do not appear to support separate justice responses based upon the categorical distinctions of adult or youth - except in the case of prison facilities. All of this leaves one wondering if the public distinguishes between justice responses for different ages of offenders at all.

Comparing youth and adults - goals of sentencing

However, a further examination of these general beliefs reveals that harsh approaches for dealing with young offenders are extensive, but are more complicated than may seem at first glance (Sprott 1998). In actually deciding what is important for different kinds of offenders, it appears that the public does distinguish, to some extent, between youths and adults.

Respondents were asked questions about the purposes of sentencing. Since half of those surveyed were responding to the adult survey and other half were asked to think about youths, this allowed me to make some comparisons. In responding to the importance of the different purposes of sentencing; expressing disapproval, deterring offenders, incapacitation, rehabilitation and providing compensation, the results show that
there were significant differences in ratings of the overall importance of different sentencing purposes. Furthermore, a comparison of ratings between youths and adults (Table 6) shows that the public does rate certain purposes as having more importance for youths than for adults.

Overall respondents rated all of the purposes of sentencing as quite important, however deterrence was rated to be the most important purpose of sentencing overall for both youths and adults (mean scores of 8.18 and 8.12). The fact that deterrence was rated as highest in importance to the public is quite interesting and reaffirms the intuitive appeal of deterrence based approaches to sentencing. Evidence of the importance of deterrence in sentencing hearings was also seen in Chapter 3, in which references to deterrence were the most frequently cited in the observed youth court hearings.

Table 6 - Comparison of public's ratings of the importance of different goals of sentencing as a function of whether the offender is a youth or an adult

<table>
<thead>
<tr>
<th>Purpose</th>
<th>youth</th>
<th>adult</th>
</tr>
</thead>
<tbody>
<tr>
<td>expressing the community’s disapproval of the crime</td>
<td></td>
<td></td>
</tr>
<tr>
<td>deterring the offender and other persons from committing offences</td>
<td>8.18</td>
<td>8.12</td>
</tr>
<tr>
<td>separating offenders from society</td>
<td>6.17</td>
<td>6.99</td>
</tr>
<tr>
<td>assisting in the rehabilitation of offenders</td>
<td>8.09</td>
<td>7.75</td>
</tr>
<tr>
<td>compensating victims or the community</td>
<td>7.57</td>
<td>7.60</td>
</tr>
</tbody>
</table>

Rated on a scale 1=not at all important 10=very important purpose.
Expressed as a mean score out of 10 for each goal of sentencing.

NOTE: main effects 'survey' (youth or adult) ; F=1, df=1, 934, p=.834 (n.s.)
main effects 'purposes' (5 purposes) ; F=79.32, df=4, 3736, p<.001
interaction effects 'survey' by 'purposes'; F=12.75, df=4, 3736, p<.001
As Table 6 shows, expressing disapproval for the crime ($r=-2.138$) and rehabilitating the offender ($r=-2.387$) were rated as having greater importance for youths than for adults. Separating offenders from the rest of society was rated as less important for youths than it was for adults ($r=5.318$). The mean differences in ratings for deterrence and compensating the victim or the community did not vary significantly in ratings for youths or adults.

What this clearly shows then, is that despite the broad beliefs mentioned earlier which suggest that the majority of the public wants a harsh approach to youths which is similar to the approach given to adults (with the exception of separating offenders), that in fact, there are variations in the public’s ratings of what sentences should be accomplishing for youthful vs. adult offenders.

**Views of the purposes of sentencing by other beliefs about the justice system:**

It may be the case, however, that public ratings of the goals of sentencing vary depending upon views of other aspects of the justice system. For instance, it is useful to examine the responses on purposes of sentencing for those that believe there should be a separate system of justice compared to those that wish to abolish a separate youth justice system. Presumably, by stating opposition to a separate youth justice system, respondents are implying that justice responses should not be based on distinctions of age.

Table 7 shows respondents’ ratings of the different goals of sentencing by views about a separate justice system.
Table 7. - Ratings of the importance of each of the goals of sentencing by support or opposition to a separate youth justice system as a function of whether the offender is a youth or an adult

<table>
<thead>
<tr>
<th>purpose</th>
<th>favour/strongly favour separate youth justice system</th>
<th>oppose/strongly oppose a separate youth justice system</th>
</tr>
</thead>
<tbody>
<tr>
<td>expressing the community's disapproval of the crime</td>
<td>youth 7.86</td>
<td>adult 7.13</td>
</tr>
<tr>
<td>deterring the offender and other persons from committing offences</td>
<td>youth 8.10</td>
<td>adult 8.07</td>
</tr>
<tr>
<td>separating offenders from society</td>
<td>youth 5.69</td>
<td>adult 6.22</td>
</tr>
<tr>
<td>assisting in the rehabilitation of offenders</td>
<td>youth 8.26</td>
<td>adult 7.78</td>
</tr>
<tr>
<td>compensating victims or the community</td>
<td>youth 7.30</td>
<td>adult 7.15</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>BETWEEN SUBJECTS</th>
<th>Effect</th>
<th>DF</th>
<th>MS</th>
<th>F</th>
<th>sig of F</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Survey</td>
<td>1</td>
<td>2.95</td>
<td>.30</td>
<td>.587</td>
</tr>
<tr>
<td></td>
<td>favour/oppose</td>
<td>1</td>
<td>126.66</td>
<td>12.66</td>
<td>&lt;.001</td>
</tr>
<tr>
<td></td>
<td>survey x favour/oppose</td>
<td>1</td>
<td>14.26</td>
<td>1.43</td>
<td>.233</td>
</tr>
<tr>
<td></td>
<td>error(between)</td>
<td>870</td>
<td>10.01</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>WITHIN SUBJECTS</th>
<th>Effect</th>
<th>DF</th>
<th>MS</th>
<th>F</th>
<th>sig of F</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Purposes</td>
<td>4</td>
<td>355.14</td>
<td>85.15</td>
<td>&lt;.001</td>
</tr>
<tr>
<td></td>
<td>survey by purposes</td>
<td>4</td>
<td>43.52</td>
<td>10.43</td>
<td>&lt;.001</td>
</tr>
<tr>
<td></td>
<td>favour/oppose x purposes</td>
<td>4</td>
<td>43.96</td>
<td>10.54</td>
<td>&lt;.001</td>
</tr>
<tr>
<td></td>
<td>survey by favour/oppose x purposes</td>
<td>4</td>
<td>2.55</td>
<td>.61</td>
<td>.655</td>
</tr>
<tr>
<td></td>
<td>error(within)</td>
<td>3480</td>
<td>4.17</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

As noted in the previous set of tables, there was an interaction of the purposes of sentencing based upon whether the offender was said to be a youth or an adult.

It appears to be the case that the two groups (those that favoured and those that opposed a separate justice system) differed from each other in terms of how they rated the various goals of sentencing. This however, is not surprising, since it is quite likely that
these two separate groups simply have different views on the relative importance of different purposes of sentencing. But most interestingly for this analysis, Table 7 also reveals that the differences in ratings on the goals of sentencing did not depend on respondent’s favouring or opposing a separate justice system. This finding has important political implications, since, on the face of it, one may be inclined to interpret the public’s expressed opposition to a separate youth justice system as evidence that they do not distinguish between offenders based upon age. However, this finding suggests, that in fact, even those that expressed opposition to a separate youth justice system still differentiated between youths and adults on the goals of sentencing in the same manner as those that favoured a separate youth justice system.

**Views of purposes of sentencing by beliefs about severity in sentencing:**

In order to see if the public’s ratings on the purposes of sentencing depended on other beliefs about the system, an analysis was completed on the purposes of sentencing by those that felt youth court sentences should be ‘harsher/the same as’ adults, and those that thought youth court sentences should be ‘less harsh than adults.’ Again, one would expect that those that expressed that youth court sentences should be ‘harsher/the same as’ adult sentences for similar kinds of offences, would probably see the purposes of sentencing for adults and youths as the same.
Table 8 - Ratings of the importance of each of the goals of sentencing by responses to whether or not youth court sentences should be harsher/the same as adults or less harsh than adult sentences

<table>
<thead>
<tr>
<th>Purpose</th>
<th>Youth</th>
<th>Adult</th>
<th>Youth</th>
<th>Adult</th>
</tr>
</thead>
<tbody>
<tr>
<td>expressing the community's disapproval of the crime</td>
<td>7.55</td>
<td>7.18</td>
<td>7.77</td>
<td>7.45</td>
</tr>
<tr>
<td>deterring the offender and other persons from committing offenses</td>
<td>8.11</td>
<td>7.98</td>
<td>8.22</td>
<td>8.28</td>
</tr>
<tr>
<td>separating offenders from society</td>
<td>5.76</td>
<td>6.36</td>
<td>6.38</td>
<td>7.35</td>
</tr>
<tr>
<td>assisting in the rehabilitation of offenders</td>
<td>8.15</td>
<td>7.88</td>
<td>8.11</td>
<td>7.72</td>
</tr>
<tr>
<td>compensating victims or the community</td>
<td>7.05</td>
<td>7.18</td>
<td>7.84</td>
<td>7.87</td>
</tr>
</tbody>
</table>

As in the previous table (Table 7), Table 8 reveals that while the two groups in this analysis (less harsh vs. harsher/the same as) differed in how they rated the various purposes of sentencing, the variation in purposes of sentencing for youths or adults did not depend upon their beliefs about the severity of sentences. Thus, purposes of sentencing for adults and youth were rated differently regardless of respondents’ views of the severity of sentencing for youths and adults.
What all of this implies then, is that even for respondents who oppose the separation of systems of justice based upon age or those that want sentences for youths to be the same as, or harsher than adults, distinctions were still made in what sentencing should accomplish for youths as opposed to adults. This may in part be explained by the finding that public punitiveness (as measured by opposition to a separate youth justice system) may be related less to a desire to punish, and more to the perception that sanctions other than prison are being ineffectively administered (Sprott 1998). Thus, it is not surprising that on a philosophical level people do distinguish justice responses based upon the age of the offender. But on a practical level, opposing differences in justice responses based upon age may be linked to other more complex issues related to the administration of justice.

**Conclusions:**

In the final analysis, the ‘youthfulness’ of a young offender appears not to influence public ratings of the offender or their preferred sentences. Instead, any descriptive information at all was associated with less harsh responses in this study. Thus, as with the two court observation chapters which showed that age and youthfulness did not affect the outcomes of court cases, the construct of ‘maturity’ also appears to have little bearing on public sentencing preferences and ratings of a young offender. Where age appears to come into the equation, however, is with respect to the purposes of sentencing for adult vs. youthful offenders. Even for those respondents who opposed a separate youth justice system or those who thought that sentences for youths should be the same as or harsher than adults, there were significant differences in what a sentence should
accomplish based on whether the offender was an adult or a youth. For the most part, the public favoured rehabilitative purposes of sentencing as being more important for youths than for adults.
Chapter Five - Conclusion

The ambivalence found in the YOA probably reflects a level of societal ambivalence in Canada about the appropriate response to young offenders. On the one hand, there is a feeling that adolescents who violate the criminal law need help to enable them to grow into productive, law-abiding citizens. ...On the other hand, there is a widespread public concern about the need to control youthful criminality and protect society (Bala 1992: 32).

The passage above notes that there is a dual approach when constructing young people in conflict with the law. There is a concern with both protecting ‘youthful offenders’ and responsibilizing ‘young offenders’. The apparent ambivalence regarding the place of youth on the ‘continuum of responsibility’ provided my initial interest in undertaking this research. Each of the preceding chapters analyzed whether decisions made about youth in conflict with the law were based upon a view of adolescence as a homogenous stage or as a developmental transition toward adulthood. The lack of clarity can, perhaps, best be understood by examining how ‘youthfulness’ is differentially regulated in other arenas depending upon the context.

For example, as noted in the introductory chapter, adolescents are prohibited from engaging in a range of ‘adult-like’ activities in wider society. And there is a fair amount of consensus among adults regarding the age at which adolescents should be able to participate in certain ‘adult-like’ behaviours (Paglia and Room 1998, Dekovic, Noom and Meeus 1997). Young people under fourteen cannot consent to sexual relations. Those who are less than sixteen in Ontario cannot receive welfare benefits. While sixteen and seventeen year olds may make a case for receiving welfare, their cheques are issued to a ‘responsible person’ on their behalf. Anyone under nineteen in Ontario cannot purchase or consume alcohol, though consumption is allowed at a younger age if served at home.
by parents. Thus, it appears from these examples that in defining when young people are
legally permitted to engage in a variety of 'adult-like' behaviours, adolescence is viewed
as a continuum where, generally speaking, a higher age is associated with fewer
prohibitions.

When it comes to the youth justice system, there is also a belief that age and one’s
level of maturity are important factors in responding to youthful criminality. The youth
justice system tries to respond to the needs of youth in conflict with the law, and also
attempts to ascertain some level of accountability for criminal activity. Section 3 of the
YOA, the Declaration of Principle, discusses young people’s state of dependency along
with the protection of the public and accountability. This applies to all young offenders
regardless of age, lending credence to the model of youth as a homogenous group. At the
same time, the YOA makes distinctions based upon chronological age specifically as it
concerns serious offences and the applicability of transfer to adult court. Therefore,
young people under the age of twelve who commit offences are not held criminally
responsible. Offenders under the age of fourteen are not eligible to be transferred to adult
court no matter how serious the offence. Young people who commit certain serious
violent offences are presumptively seen as adults if they are sixteen or seventeen years
old. And at eighteen, criminal accountability occurs in the adult criminal justice system.

Therefore, the view of adolescence in the youth justice context is ambiguous.
There appear to be two different models constructing adolescence at work in the youth
justice context. Adolescence can be interpreted either as a distinct stage or as a period of
developmental transition leading up to adulthood.
The ambivalence in responding to youth who commit offences has been apparent since the inception of separate legislation. Youth who are fourteen and older have always been subject to transfer to adult court —under both the JDA and the YOA. In the original YOA there was a short-lived provision which stated that secure custody could not normally be used for those under fourteen. In addition, Bill C-37 (1995) created a further demarcation with the presumptive transfer provisions at the ages of sixteen and seventeen. So within the framework of separate youth justice legislation in Canada since 1908, there have always been stipulations attached to individual cases, where chronological age becomes important.

How do these models play out in the youth justice system?

An interesting example of the differing interpretations of mitigated responsibility comes from the Supreme Court of Canada case R v. J (T.J). The young offender in this case was a mature seventeen year old charged with first-degree murder in the sexual assault and murder of a three-year old girl. He had a common-law wife, a child, and also worked for his cousin as a roofer. At issue was whether or not the police complied with the Charter of Rights and Freedoms as well as s.56 of the YOA, which provides additional protections to young people at arrest in terms of questioning, and the taking of statements. In trying to understand how a youth’s maturity and competence factors into their ability to understand their rights, Justice Cory put forward the following:

By its enactment of s. 56, Parliament has recognized the problems and difficulties that beset young people when confronted by authority. It may seem unnecessary and frustrating to the police and society that a worldly-wise smug, 17-year-old with apparent anti-social tendencies should receive the benefit of this section. Yet it must be remembered that the section is to protect all young people of 17 years or less (R v. J(T.J)(1990) 59 C.C.C.(3d) 1 (S.C.C.)p.8).
L’Heureux-Dubé J. takes quite a different stance in accounting for the effects of maturity on decision-making. After referring to arguments made by Parliamentarians who were opposed to the inclusion of sixteen and seventeen year olds into the YOA, she goes on to relate the following:

...young offenders suspected of criminal offences should be treated in a manner befitting their ages. Adolescence cannot be viewed as a snapshot in time. Those youths between the ages of 12 and 18 cannot be aggregated and dealt with uniformly without regard for the discrepancies in their faculties and competence.

Within this “child-adult grouping” there are those that are more “child” and those that are more “adult”. We should be especially sensitive to 12- and 13-year olds at the younger end of the spectrum. Their youth borders on that age considered too young to be included within the scope of the Act entirely. Conversely, 17-year-olds are on the brink of adulthood, months away from attaining their full measure of protection under the Canadian Charter of Rights and Freedoms, but no more(R v. J(T.J)(1990) 59 C.C.C.(3d)1 (S.C.C.)p.14,15).

Even at the level of the Supreme Court, there is disagreement in interpreting how an offender’s age and state of maturity apply to youth court cases. The findings from the present study provide some degree of insight into the practical application of age related constructs in youth justice decision-making.

Summary of Findings

The two court observation studies and the analysis of the Statistics Canada data support the perspective that youth in conflict with the law are treated as a homogenous group. Though the maturity of a youth might seem as if it should be important, age (both chronological and apparent) does not appear to be important in decision-making. The results from this study indicate that at two major decision making points in the youth justice process (pre-trial detention and sentencing) young people, for the most part, are dealt with as a homogeneous category. This is interesting because at one point (their 12th birthday) they are one day beyond being incapable of being criminally responsible, and
the day before their 18th birthday they are one day before being fully responsible for their crimes as adults. Only on some of the more marginal issues like 'boundary conditions' and 'curfew orders' is there any indication that age is a consideration in court decisions.

The public, as well, seems not to respond to the apparent (social) maturity of youths, but does differentiate in how they want young people to be sentenced. Most notably, the public rated different goals of sentencing to be important depending upon whether the offender was an adult or a youth. This is not a "leniency-harsness" finding. Instead, there appears to be an interest in distinguishing correctional responses for youthful offenders compared with adults.

The Disappearance of Youth?

What this may suggest is that for criminal justice purposes, the public and the youth justice system find it is easier to think of youth who are between twelve and eighteen as a single homogenous group. Incorporating the differences among youth demands a further dimension of decision-making that is not currently being considered by court practitioners. For example, this study has found that court decisions are based mainly on legal factors, as is the case with adults. But there is little in the way of differentiating among youth unless one gets down to the level of individual controls. For example, only in cases of serious violent offences is there explicit mention of an offender’s chronological age (in reference to transfer). And as revealed in the present study, in less significant areas, such as with boundary conditions and curfews, ‘younger’ youth were treated differently --they were more likely to be given these conditions than were ‘older’ youth.
It is also useful to remember how this homogenized view of youth is characterized. For instance, there was little evidence in this study to suggest that youth were being treated explicitly as ‘youth’ in court. There was no mention of the ‘responsible person’ provision specific to youth in pre-trial detention hearings. And there was limited use of the special sentencing principles of the YOA with the exception of the ‘adult’ principle of deterrence. Therefore the homogenized view of youth appears to be more closely associated with the principles of the adult justice model. As a consequence, young people at one end of the spectrum (12 and 13 year olds) are being grouped together and dealt with primarily within a deterrence-based model aimed at controlling the offending behaviour of older youth.

**Remembering that there are differences among youth:**

The differences that exist among youth should be carefully considered. As noted in the introductory chapter, a large body of developmental research informs us of the significant differences among adolescents in this age group. Younger youths are at a greater disadvantage when it comes to understanding and competently participating in the legal process (Scott and Grisso 1997, Abramovitch, Higgins-Biss and Biss 1993, Peterson-Badali and Koegl 1999). Thus, it is surprising to see that court decisions do not vary based upon age group. Furthermore, it is disconcerting to note that the adversarial process for youth in courts is limited. This study showed that concrete case planning by defense counsel in sentencing cases was a rarity. In addition, the decision to detain at bail hearings was predominantly a function of the decision made by the Crown Attorney. Not only does this crime control atmosphere put all young offenders at a disadvantage, but
given the greater lack of understanding among younger youth, the protection of their due process rights are further compromised.

Developmental research also shows that differences exist among adolescents in terms of rational choice and reasoning skills. Younger adolescents engage in simpler decision-making processes. They are less capable of imagining risky consequences, and only consider a limited range of consequences during hypothetical problem solving situations (Scott and Grisso 1997). Given these differences, it is also surprising to find that when references to principles of sentencing were made in court in this study, the most predominant was the principle of deterrence. The presumption that youth rationally choose to commit offences is not supported by the available evidence (Doob, Marinos and Varma 1995).

"Re-establishing" youth as a concept:

To some extent the findings reported here are not surprising since the YOA resembles the adult justice model in many respects. However, the recently proposed replacement to the YOA, Bill C-3 (1999), the Youth Criminal Justice Act may represent an attempt at 're-establishing' the category of youth as more differentiated from the adult justice model than is the case with the YOA. The YCJA puts forward special procedures for youth in conflict with the law that do not exist for adults. For example, at various stages in the justice process, the YCJA supports the use of extrajudicial measures. At the initial stage of the process, police officers are encouraged to refer youth to community-based programs, use cautions, or take no measures at all where appropriate. There are also explicit provisions at other stages in the process. For pretrial detention hearings, a
judge must inquire into the availability of a responsible person, which presumably will lead to the greater use of this provision. In addition, detention can only be used in cases where the offence would, upon conviction, warrant a committal to custody. In sentencing, there are specific principles and factors to be considered in sentencing youths. There are also explicit restrictions on the use of custody, reserving its use for violent offenders or those that have previously failed to comply with non-custodial sentences.

The YCJA appears to be more explicit in creating leniency or special treatment for youths as opposed to adults which stands in stark contrast to the selective incapacitation rationale being used in sentencing in the state of Virginia. In Virginia, youthful status has become an *aggravating* factor at sentencing for drug offences, fraud and larceny. The younger the offender, the more points there are against him and the more likely he will be placed in custody (Tonry 1999).

While the newly proposed YCJA appears to carve out a more distinct space for youth in conflict with the law, there are still some significant challenges to ‘re-establishing youth’. First, the present study reveals that existing legislative provisions are not necessarily utilized in court decisions. This analysis has revealed that there was little use made of the specific provisions of the *YOA* in court hearings. While this may be a function of the general similarities between the *YOA* and the adult justice model, it still suggests that changes in legislation may not enter the realm of youth court decision-making and that there is very little of ‘seeing youth’ taking place in court hearings. Consequentially, the likelihood that new legislation will fundamentally alter decision-making practices is evidently problematic.
Second, there is still the persistent issue of ambivalence surrounding the treatment of youth. For example, the category of ‘youth’ appears to be even more homogenized under the *YCJA*. As recently as 1994, Bill C-37 specified differences among youth in terms of transfer, where sixteen and seventeen year old offenders would be presumptively transferred for serious violent offences, but the general age at which transfer was applicable was fourteen. Only six years later, the *YCJA* proposes to lower the age of presumptive transfer to include fourteen and fifteen year old youths. Now, only twelve and thirteen year olds are explicitly distinguished from the rest of youth in the proposed legislation. So while there appears to be a move to re-establish youth as a concept in the form of special procedures apart from adults, there does not appear to be a movement to establish differences among youth.

Therefore, the *Youth Criminal Justice Act* may end up carving out a space for youth which is either broad or deep. A deepening of the space may accommodate the wide range of developmental differences that this six-year stage of development encompasses. The greater use of extra-judicial measures and special procedures which are distinct from the adult model may allow for more meaningful responses, thus distinguishing among young people coming before the law. The more distressing view is that this space may be broadened, and the categories of ‘youth’ collapsed even further as can be seen with the proposal to lower presumptive transfer to the age of fourteen.

Additionally, if the lack of consideration for youth-centred principles continues to exist under the new legislation, this may result in casting the net even wider. For example, as this research has shown, when judges did take account of youthful status, it appeared to be in relation to the use of ‘boundary conditions’ or ‘curfew’ orders for younger youth.
Thus, while trying to account for youthful status, the tools that judges use to do this may result in a widening of the net of control and the potential for greater coercion of younger youths. Thus, the courts may continue to decide cases involving very young adolescents in a manner which resembles the adult criminal model, and in accounting for ‘youth’ the result may be increased social control.

Conclusion

This study provides a first step in understanding how decisions about youth in conflict with the law are affected by notions of age and maturity. But, in view of the fact that the court observation data were collected only in Toronto, additional research on decision-making in courts in other locations would be highly useful. It may be possible that significant relationships exist between age and court outcomes in smaller communities, where court practitioners may have a greater knowledge of the background of particular youths. It would also be valuable to understand the factors that relate to court decisions made for adults and how these compare to the findings reported for youth in this study.

What we can take from this study is a greater understanding of the construction of youth in conflict with the law. If we are concerned with carving out a space for young offenders, this space must be able to accommodate the developmental differences of this population. Furthermore, the specific provisions of the legislation need to be reflected in courtroom decisions. The proposed Youth Criminal Justice Act may provide the means to improve our response to all youth who come in conflict with the law. Future research will tell us if we are successful in doing so.
References


228


INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

Bell & Howell Information and Learning
300 North Zeeb Road, Ann Arbor, MI 48106-1346 USA
800-521-0600

UMI®
ON THE DISCRETE LOGARITHM PROBLEM IN FINITEIELDS AND ON ELLIPTIC CURVES

by

Theodoulos Garefalakis

A thesis submitted in conformity with the requirements for the degree of Doctor of Philosophy Graduate Department of Computer Science University of Toronto

Copyright © 2000 by Theodoulos Garefalakis
The author has granted a non-exclusive licence allowing the National Library of Canada to reproduce, loan, distribute or sell copies of this thesis in microform, paper or electronic formats.

The author retains ownership of the copyright in this thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without the author’s permission.

0-612-53692-0
Abstract

On the discrete logarithm problem in finite fields and on elliptic curves

Theodoulos Garefalakis
Doctor of Philosophy
Graduate Department of Computer Science
University of Toronto
2000

The discrete logarithm problem has been the basis of numerous cryptographic schemes. The security of those cryptosystems is based on the presumed intractability of the discrete logarithm problem in some group. The subject of this thesis is the discrete logarithm problem in the multiplicative group of finite fields and in the group of points of elliptic curves over finite fields.

The most successful algorithm for computing discrete logarithms in finite fields is the index calculus method. A significant parameter of the method is the factor base, which has invariably been chosen to contain all irreducible elements up to some size. Until now, this choice had not been questioned. We attempt a rigorous study of the behavior of the algorithm in the case that a different factor base is used. This leads us to generalize the notion of smoothness, and prove density theorems for those generalized smooth polynomials analogous to the existing ones for smooth polynomials. Subsequently, we use them to analyze the index calculus method, that operates with a non-smooth factor base. The analysis shows that the more general version of the method has the same asymptotic running time, as the traditional version.

For the discrete logarithm problem on elliptic curves, our goal is to obtain a polynomial time reduction to the discrete logarithm problem in finite fields. A complete such reduction has been given by Frey and Rück. Our reduction is a generalization of that of Menezes, Okamoto, and Vanstone. We present a generalization of the Weil pairing, that is parameterized by an isogeny $\psi$. Then we specialize the isogeny to $1 - \phi$, where $\phi$ is
the Frobenius endomorphism, and use it to construct an isomorphism between the group of points of interest and a suitable group of roots of unity. For an elliptic curve $E/\mathbb{F}_q$, and a point $P$ in $E(\mathbb{F}_q)$ of prime order $r$, our construction works for $r|q - 1$. Finally, we present an efficient algorithm for evaluating the pairing, which makes the isomorphism efficiently computable. Our contribution is a new, conceptually simpler construction, that is equivalent to the one by Frey and Rück. The Frobenius endomorphism has been used in the past to speed up the arithmetic on elliptic curves, as well as to speed up Pollard’s $\rho$ method for computing discrete logarithms.
Acknowledgements

This thesis would not have been completed without the support and encouragement of several people. The first persons to thank are my supervisors, Allan Borodin and Daniel Panario. Allan has always been ready to help. His remarkable perspective and experience are unmatched, and this thesis has benefited from them. Daniel has been the person that guided my research and encouraged me during the past three and a half years. I thank him for sharing with me his experience and his enthusiasm. Being Allan’s and Daniel’s student is a privilege.

I am deeply grateful to the members of my committee: Charlie Rackoff, Ian Blake, Kumar Murty, John Friedlander, and Nigel Boston. I feel extremely fortunate that I had the opportunity to discuss and learn from each one of them. The interaction with such a group of distinguished scientists (or maybe artists) is a rare and most valuable educational experience.

Several people made the past five years in Toronto enjoyable. Panayiotis Tsaparas, my house mate for all those years was always happy to hear my problems and concerns, and gave me his unique point of view. Fotis has also been a good house mate and friend. The Faloutsos brothers, Petros and Michalis, have guided me since the beginning, and have been great friends to me. Bil Tzerpos, probably the most relaxed person I know, has been a good teacher of pool, and a good student of bridge. He and Mel have organized the best home-gatherings ever. Nick Koudas continues to be the best source of good jokes, and the best person to discuss history and politics. I am sure that his latest readings have made him wiser. Lucia Moura, my Brazilian friend has always offered encouraging words and a smile. I will never forget her cooking skills, and the parties she organized together with Daniel.

I want to thank Themis Palpanas, Yiannis Velegrakis and Rozalia Christodoulopoulou for their smile, and Yiannis Papoutsakis for being a good friend for all those years. A special “thank you” goes to Periklis Andritsos, who has been the best of friends, a person I always knew I could depend on. Many thanks to my former office mates Stergios Anastasiadis, and Tasos Kementsietsidis for offering me their help whenever I needed it. Their practical minds saved me a lot of unnecessary work.

Things would have have been harder during my graduate studies if it wasn’t for Kathy Yen, our graduate secretary. I thank her for all her efforts to make our lives easier, but primarily I thank her for being a friend.

I want to thank Irini Dimou, my partner, for all her support and encouragement. I feel very lucky to be with a person of her intelligence, honesty, and warm personality.
am grateful to her for making my life complete.

Last but certainly not least, my utmost thanks to my parents Litsa and Stelios, and my sister Annabela for teaching me things that cannot be found in any university curriculum. Their love and support made this thesis possible. It is dedicated to them with all my love and appreciation.
Contents

1 Introduction .......................... 1
  1.1 The discrete logarithm problem ...................... 1
  1.2 Generic algorithms ................................ 3
  1.3 The discrete logarithm problem in finite fields ....... 4
    1.3.1 Rigorous subexponential algorithms ............... 5
    1.3.2 Heuristic subexponential algorithms ............... 5
  1.4 The discrete logarithm problem on elliptic curves .... 6
  1.5 Overview of the thesis ................................ 8

2 Preliminary topics ................... 13
  2.1 Algebraic preliminaries .......................... 13
    2.1.1 Finite fields .................................. 14
  2.2 Analytic and combinatorial techniques ............... 16
    2.2.1 Generating functions .......................... 16
    2.2.2 Convergent power series ....................... 18

3 Density theorems .................... 21
  3.1 Motivation and results .......................... 21
  3.2 A technical lemma ................................ 22
  3.3 Fixed lower bound ................................ 25
  3.4 Lower bound tending to infinity .................... 27

4 The discrete logarithm problem in finite fields ........ 37
  4.1 Introduction ................................... 37
  4.2 The algorithm .................................. 38
  4.3 Analysis ....................................... 39
    4.3.1 Size of the factor base ....................... 40
    4.3.2 Probability of success ....................... 40
4.3.3 Time complexity ................................................. 42
4.4 Some remarks ...................................................... 45
4.4.1 The Waterloo variant ........................................... 45
4.4.2 The Coppersmith variant ....................................... 46
4.5 Experimental results ............................................... 47

5 Basic notions on elliptic curves 49
5.1 Elliptic curves ....................................................... 49
5.2 Polynomial and rational functions ................................ 50
5.3 Zeros and poles ..................................................... 52
5.4 Divisors .............................................................. 54
5.5 The group law ........................................................ 55
5.6 Lines ................................................................. 57
5.7 Maps between elliptic curves .................................... 60
5.8 Isogenies .............................................................. 62
5.9 The dual isogeny ..................................................... 65
5.10 Weil reciprocity ..................................................... 66

6 The elliptic curve discrete logarithm problem 69
6.1 Introduction ......................................................... 69
6.2 Reductions to finite fields ....................................... 70
  6.2.1 The MOV reduction ........................................... 71
  6.2.2 The FR reduction ............................................. 73
6.3 The generalized Weil pairing .................................... 74
6.4 Construction of new isomorphisms ............................... 79
  6.4.1 Curves with \( a_q = 2 \) ....................................... 80
  6.4.2 A structure theorem .......................................... 81
  6.4.3 Computing the pairing ....................................... 84
6.5 Generalizations .................................................... 87
  6.5.1 The case \( r | q - 1 \) .......................................... 88
  6.5.2 The case \( r | q^b - 1 \) ........................................ 89

7 Conclusions ........................................................... 95
List of Figures

3.1 Decomposition of the integral ............................................. 30
4.1 The form of the factor base .............................................. 44
List of Tables

4.1 $n = 500$, $p = 2$, $m_1 = 47$ .......................... 48
4.2 $n = 500$, $p = 3$, $m_1 = 38$ .......................... 48
4.3 $n = 500$, $p = 5$, $m_1 = 31$ .......................... 48

6.1 Reductions of the ECDLP to DLP in finite fields .......... 74
Chapter 1

Introduction

1.1 The discrete logarithm problem

Cryptography, the art of secret writing, has grown into a field of research by its own. Its methods and definitions are precise and exist within the framework of complexity theory, a well established field in the intersection of computer science and mathematics. Thus the development of cryptographic primitives is a process based on mathematical problems. and the "security" of those primitives is directly related to the properties of the underlying problems.

An obvious requirement of a good cryptographic system is that messages should be easy to encrypt and decrypt for legitimate users, and at least decryption should be hard for everyone else. Number theory has turned out to be an excellent source of computational problems that have both easy and apparently hard aspects, and can be used as the basis for several cryptosystems. Examples of such problems are integer factorization, the discrete logarithm problem in several groups, and lattice basis reduction problems.

This thesis is concerned with some computational aspects of the Discrete Logarithm Problem (DLP). The Discrete Logarithm Problem has been the basis of several cryptographic schemes since 1976, when Diffie and Hellman in their seminal paper [17] proposed the first public-key cryptographic protocol for key exchange. Their construction was based on the assumption that a certain problem, now known as the Diffie-Hellman problem (DH), is computationally intractable. Since then, the ideas of Diffie and Hellman have been extended in many directions, and several cryptographic schemes have been developed, that solve different problems and enjoy different algebraic properties. However different, all those cryptosystems base their security on the difficulty of Diffie-
Hellman problem for some group $G$. In the following definition, we write the group $G$ multiplicatively.

**Definition 1.1.1 (Diffie-Hellman Problem)** Given a finite abelian group $G$, an element $g \in G$, $g^a$ and $g^b$, for some $a, b \in \{1, \ldots, \#(g)\}$, compute $g^{ab}$.

It is clear that the DH problem can be easily solved, if one can compute $a$, given $g$ and $g^a$. This is the DLP for the group $G$.

**Definition 1.1.2 (Discrete Logarithm Problem)** Given a finite abelian group $G$, an element $g \in G$ and $y \in \langle g \rangle$ compute the smallest positive integer $\ell$ such that $y = g^\ell$. The integer $\ell$ is called the discrete logarithm of $y$ to the base $g$ and is denoted by $\log_g y$.

In cryptography, one is interested in the average case complexity of the above problems. By that we mean that the integers $a$ and $b$ in the DH problem and the element $y$ in the DLP are chosen uniformly at random (in the corresponding domains). For the DLP, the worst case complexity is the same as the average case complexity: given any $y \in \langle g \rangle$, one can choose $s \in \{1, \ldots, \#(g)\}$ uniformly at random, and compute $y g^s$. Thus the worst case is reduced to the average case.

The DH problem is trivially reduced to the DLP in the same group. The original DH assumption asserts that the DH problem is intractable for the multiplicative group of the integers modulo a large prime. Until today, the original DH assumption remains an open problem. Surprisingly, the only method known to solve the DH problem is in fact to solve the DLP first. Little progress has been made on the DH problem. Most results are contained in [40, 41], where some connections to the DLP are proven.

On the other hand, significant progress has been made on the DLP. Before discussing the progress on the DLP, the following observation is in order. The DLP as defined is really a problem in cyclic groups (that are necessarily abelian). All cyclic groups of given order, say $n$, are isomorphic to the additive group $\mathbb{Z}/n\mathbb{Z}$, where the DLP is trivial. However, as we already mentioned, the DLP in $(\mathbb{Z}/p\mathbb{Z})^*$ remains intractable. The reason is that no efficient algorithm is known to compute the isomorphism to the additive group. This indicates that the difficulty is not group theoretic — group theoretically there is one cyclic group. It is rather the way that the elements of the group are represented that makes the problem easy or hard. In view of this remark, we move on to discuss three situations: when no information is given about the group (other than its order), when it is the multiplicative group of a finite field, and when it is the (additive) group of points on an elliptic curve over a finite field.
1.2 Generic algorithms

We start our exposition with a review of the basic results regarding *generic algorithms*. They are algorithms for the DLP that do not exploit any special properties of the group at hand, other than that each group element is encoded by a unique binary string. One can imagine that such an algorithm has access to the group only through a black box, that performs the basic group operation, inversion, and identity test. For a formal definition of generic algorithms, the reader is referred to the paper by Shoup [63].

The first non-trivial generic algorithm for computing discrete logarithms and related problems, the *Baby Step Giant Step* method, was given by Shanks in [62]. The basic idea is as follows. We are given \( g \in G \) and \( y \in \langle g \rangle \), and we want to compute the integer \( x \in \{0,\ldots,n-1\} \) that satisfies \( y = g^x \), where \( n = \#\langle g \rangle \). Let \( m = \lceil \sqrt{n} \rceil \), and write \( x = im + j \), with \( 0 \leq j < m \). Such integers \( i, j \) always exist by long division. Then we also have \( 0 \leq i < m \), since otherwise we would have \( x > n \). Now we have

\[
y = g^{im+j} \iff g^{-j}y = g^{im}.
\]

We first compute \( g^{im} = (g^m)^i \) for \( i = 0,\ldots,m-1 \) (the giant steps) and store them in a sorted array. Then we start computing the values \( g^{-j}y, j = 0,\ldots,m-1 \) (baby steps) one by one and compare them with the stored values. When a match is found we are done. The algorithm takes time \( O((\log n)^e \sqrt{n}) \), where the \( (\log n)^e \) factor comes from sorting and searching the array.

The next significant advance appeared in [51], where Pohlig and Hellman described a method that solves the DLP in \( \langle g \rangle \) in time \( O((\log p)^e \sqrt{p}) \), where \( p \) is the largest prime divisor of the order \( n \) of \( g \). The idea is to compute the discrete logarithm \( x \) modulo the prime powers that divide \( n \) and then combine the results using the Chinese Remainder Theorem. The main contribution of this algorithm is that it made clear that the difficulty of computing discrete logarithms in a group \( G \) comes entirely from the subgroups of \( G \) of prime order. Consequently, if the group order \( n \) has only small prime factors, then the problem is easy (even though \( n \) itself is large). For that reason precisely, we will consider in Chapter 6 only subgroups of prime order.

The algorithms of Shanks and Pohlig and Hellman improve significantly upon the trivial brute-force search, but they remain of exponential time complexity in the input size (which is \( O(\log n) \)). One might expect that further research would yield better algorithms. However, this turned out not to be the case. In [63], Shoup proved (among other things) that \( \Omega(\sqrt{p}) \) is lower bound for the time complexity of any generic algorithm. Here \( p \) is again the largest prime divisor of the group order \( n \). Shoup's lower bound matches the known upper bounds, if we disregard the log factors.
Shoup’s results suggest that the only way to obtain improved algorithms for the discrete logarithm problem is to consider specific groups and make essential use of the encoding of the group elements; that is, one has to take advantage of the properties of the specific group at hand. In the following two sections, we review the main results regarding two groups that are widely used for cryptographic purposes.

1.3 The discrete logarithm problem in finite fields

The first group to be proposed for cryptographic purposes was the group of nonzero integers modulo a prime \( p \). This is precisely the multiplicative group of the finite field \( \mathbb{F}_p \) of \( p \) elements. Subsequently, the multiplicative group of extension fields was used; that is the multiplicative group of fields \( \mathbb{F}_q \), with \( q = p^k \) for \( p \) prime and \( k \geq 1 \). From all those fields, the prime fields \( \mathbb{F}_p \) and the finite fields of characteristic 2, \( \mathbb{F}_{2^k} \), have a prominent place, mainly due to reasons of implementation. Discrete logarithms in finite fields form the basis of several cryptographic schemes, including the Digital Signature Standard. It is therefore important to understand the complexity of the problem, in order to gain confidence on the security of the cryptosystems. Unlike the generic algorithms, however, no lower bounds are known for this case. So it becomes important to study the problem and benchmark how hard it is at the present.

By far the most successful method for computing discrete logarithms in finite fields is the index calculus method. The approach first appeared in the 1920’s, in the work of Kraitchik [34], and was rediscovered in the 1970’s by Adleman [2], Merkle [44], and Pollard [52]. For a general description of the method, let \( \mathbb{F}_q \) be a finite field, \( g \in \mathbb{F}_q \) a generator of the multiplicative group \( \mathbb{F}_q^* \), and \( S \) a set of elements of \( \mathbb{F}_q^* \). We call \( S \) the factor base. Suppose \( y \) is the element whose discrete logarithm we want to compute. The algorithm works in two stages: in the first stage we collect identities of the form

\[
\prod_{v \in S} v^{e_v} = g^s.
\]

Each such identity gives a linear congruence

\[
\sum_{v \in S} e_v \log_g(v) \equiv s \pmod{q - 1}.
\]

Once we have gathered enough congruences and the linear system has full rank, we solve it to obtain the logarithms of the elements in \( S \). In the second stage, we try to construct a relation of the form

\[
\prod_{v \in S} v^{e_v} = yg^s,
\]
which implies
\[ \sum_{v \in S} e_v \log_g(v) - s \equiv \log_g(y) \pmod{q-1}, \]
and the desired logarithm can be computed.

Since the 1970's, several variants of the index calculus have appeared in the literature. The goal in every work was either to describe a version of the method that is rigorously analyzable, or to produce a method that works well in practice at the cost of not being provable. In the next two sections, we review some of those variants.

1.3.1 Rigorous subexponential algorithms

The only rigorous subexponential algorithms known have time complexity of the form
\[ \exp\left( (c + o(1)) (\log q \log \log q)^{1/2} \right), \]
where \( q \) is the number of elements in the field, and \( c \geq 1 \) is a constant. It is convenient to use the following notation for the running time of index calculus-type methods
\[ L_{p^2}(s; c) = \exp\left( (c + o(1)) (\log q)^s (\log \log q)^{1-s} \right). \]

For field \( \mathbb{F}_q \), with \( q = p^n \) for \( p \) fixed and \( n \to \infty \) the running time of the algorithm depends on the probability that a monic polynomial of degree \( n \) has only irreducible factors of degree \( \leq m \) for some suitable \( m \). Odlyzko [48] was the first to estimate this probability for \( p = 2 \). The running time of the algorithm then follows and is \( L_q(1/2; \sqrt{2}) \). Lovorn [37], and Lovorn Bender and Pomerance [38] extended the arguments, and using also results of Soundararajan proved that the algorithm has expected running time \( L_q(1/2; \sqrt{2}) \) as long as \( p \leq n^{\Omega(1)} \) and \( q \to \infty \). Garefalakis and Panario [24] generalized the results of Odlyzko [48] and Lovorn [37], and used them in [25] to prove that the running time of a more general version of the index-calculus has time complexity of the exact same form. These results are presented in Chapters 3 and 4 of this thesis.

For prime fields \( \mathbb{F}_p \) and quadratic extensions \( \mathbb{F}_{p^2} \) subexponential algorithms are also known. For the first case, Pomerance [53] showed that the basic method works in time \( L_p(1/2; \sqrt{2}) \) for \( p \to \infty \). For the second case, Lovorn [37] modified the approach of ElGamal [18] to obtain a variant that she was able to analyze. The running time of her method is \( L_{p^2}(1/2; 3/2) \), for \( p \to \infty \).

1.3.2 Heuristic subexponential algorithms

The algorithms discussed in the previous section have been modified and extended in several ways over the past few years. What is gained is a broader range of fields for
which they apply, and significantly improved running times. What is lost is the rigorous analysis.

The first algorithm that is conjectured to run in subexponential time for any finite field was presented by Adleman and DeMarrais [4]. The conjectured running time is $L_q(1/2; c)$, for $q \to \infty$. The constant $c$ depends on the field, and is at most 2. The heuristic analysis depends on the Generalized Riemann Hypothesis (GRH) as well as other unproven hypotheses concerning smooth elements in rings of integers of number fields.

Perhaps the most dramatic development has been the invention of variants of the index calculus that are conjectured to run in time $L_q(1/3; c)$. This development started with the "Waterloo algorithm" presented by Blake et al. in [7]. While their algorithm is not quite as fast, their idea of substituting the factorization of one polynomial of degree approximately $n$ by two polynomials of smaller degree inspired Coppersmith to develop his method. The "Coppersmith method" [13] works for fields of characteristic 2, and was the first to have (heuristic) running time of the form $L_q(1/3; 1.35\ldots)$.

The generalization of the ideas of Coppersmith led to the development of the Function Field Sieve (FFS) by Adleman [3], and Adleman and Huang [5]. The FFS is designed to work for extension fields of relatively small characteristic. For $\mathbb{F}_q$, with $q = p^n$ and $n \geq (\log p)^2$ heuristic arguments suggest a running time of the form $L_q(1/3; (64/9)^{1/3})$.

For prime fields and small field extensions — i.e., when the extension degree is small compared to the characteristic of the field — the FFS performs poorly. The algorithm of choice in this case is the Number Field Sieve (NFS). In this case, the motivation came from the development of an algorithm for integer factorization, the number field sieve factoring algorithm. Gordon [27] realized that the NFS factoring algorithm could be modified to compute discrete logarithms in $\mathbb{F}_q$. Since then, several researchers have further developed the method, and made it to work in conjectured time $L_q(1/3; (64/9)^{1/3})$ for fields $\mathbb{F}_q$ with $q = p^n$ and $n < (\log p)^{1/2}$ (see [55, 56, 57]). We should mention, that although the NFS is a very recent method, a special case was described by Coppersmith, Odlyzko, and Schroeppel in [14] a decade earlier.

1.4 The discrete logarithm problem on elliptic curves

The developments outlined in the previous sections motivated mathematicians to look for alternative groups, where no subexponential time algorithms for the discrete logarithm problem are known. The most exciting such proposal involved the group of points of elliptic curves over finite fields [33, 45]. No general purpose subexponential time algorithm
i.e., an algorithm that works for every elliptic curve $E$ over any finite field $\mathbb{F}_q$ — for the discrete logarithm problem on elliptic curves (ECDLP) is currently known. In fact, many believe that no such method exists, as the index calculus does not seem to apply [65]. For that reason, the elliptic curve analogs to the “traditional” cryptosystems based on the DLP in finite fields, offer the same level of security with considerably smaller keys. This led to the development of elliptic curve cryptosystems for commercial use for applications including mobile telecommunications, data encryption tools, security protocols for electronic commerce, digital signature generation by smart cards, and many others [11].

We postpone the formal definition of elliptic curves and the group law until Chapter 5. For this discussion, it is enough to think of an elliptic curve given by the equation

$$Y^2 = X^3 + aX + b, \quad a, b \in \mathbb{F}_q, \quad (1.1)$$

where the cubic on the right has three distinct roots over $\overline{\mathbb{F}}_q$, as the set of points $(x, y) \in \mathbb{F}_q \times \mathbb{F}_q$ that satisfy Equation (1.1), together with a special point $O$. The set

$$E(\mathbb{F}_q) = \{(x, y)|y^2 = x^3 + ax + b\} \cup \{O\}$$

equipped with a point addition law, forms a group with $O$ serving as the identity. It can by proved that $\#E(\mathbb{F}_q) = q + 1 - a_q$, with $|a_q| \leq \sqrt{q}$. Thus the size of the group is close to that of the underlying finite field, but is not, in general, the same. Traditionally, the group of points of an elliptic curve is written additively. Then, the discrete logarithm problem is formulated as follows. Given an elliptic curve $E$ defined over $\mathbb{F}_q$, a point $P \in E(\mathbb{F}_q)$ of order $r$ and a point $Q \in \langle P \rangle$, find the least positive integer $\ell$ such that $Q = \ell \cdot P$.

As it is the case with traditional discrete logarithm based cryptosystems, no proofs of security are known for their elliptic curve analogs: that is no proofs of the intractability of the underlying problem is known. However, the ECDLP appears to be trickier than its finite field analog. The reason is that its difficulty depends not only on the finite field, but also on the defining equation. When the order $q + 1 - a_q$ of the group is divisible only by small primes, then the problem is of course easy, as the Pohlig-Hellman method applies. In the next paragraphs we describe some of the successful attacks to the ECDLP. They all work in special cases, but they are interesting both theoretically and practically. From a theoretical point of view, they exploit deep mathematical properties of elliptic curves and reveal connections between the ECDLP and the DLP in finite fields. From a practical point of view, these attacks indicate what classes of curves are to be avoided in practical systems.
Historically, the first subexponential algorithm was discovered by Menezes, Okamoto, and Vanstone [42]. They used the Weil pairing to reduce the discrete logarithm problem on $E(F_q)$ to the discrete logarithm problem in the multiplicative group of $F_{q'}$. The degree $l$ of the extension is in general very large (on the order of $q$), but Menezes et.al. showed that for the class of supersingular curves $l$ is at most 6. Supersingular curves can be identified by the simple condition $p|a_q$. In the case that $q$ is a prime greater than or equal to 5, then this is equivalent to $a_q = 0$. This reduction combined with a subexponential algorithm for the DLP in $F_{q'}$ yields the desired algorithm for $E(F_q)$.

Later, Frey and Rück [20] and Frey, Müller, and Rück [19] presented a different algorithm based on the Tate-Lichtenbaum pairing. The method reduces the ECDLP in $E(F_q)$ to the DLP in $F_q$ and works for curves with $a_q \equiv 2$ modulo the order of the base point (say $r = \#(P)$). This condition is equivalent to $r|q - 1$, and is clearly necessary. By the theorem of Frey et.al. it is also sufficient. We note that this method can be generalized to work in the case $r|q^l - 1$, in which case the reduction is from $E(F_q)$ to $F_{q'}$, although this was not proved in [19].

Finally, there is a third, perhaps more surprising result, which asserts that for anomalous curves, i.e., curves with $a_q = 1$, the discrete logarithm problem can be solved in polynomial time. The result was discovered independently by Smart [66], Semaev [59], and Satoh and Araki [54]. In this case an efficiently computable homomorphism is constructed between $E(F_q)$ and the additive group of $F_q$. The ECDLP is thus reduced to the DLP in the additive group of $F_q$, where it is trivial to compute.

As the work of this thesis is related to the reductions of Menezes et.al. [42] and of Frey et.al. [19], we will give further details on them in Chapter 6. For further details on “anomalous attack” as well as other subjects related to the mathematics and implementation of elliptic curve cryptosystems we refer to the book by Blake, Seroussi, and Smart [6].

1.5 Overview of the thesis

This thesis focuses on the discrete logarithm problem in finite fields, and on the group of points of elliptic curves over finite fields. Our goal in both cases, is to present algorithms that work provably in subexponential time.

The thesis is logically divided into two main parts. The first part — Chapters 2, 3, and 4 — deals with the discrete logarithm problem in finite fields. Our goal is the analysis of an algorithm for the DLP in $F_q$, which is presented in Chapter 4. The second part — Chapters 5 and 6 — focuses on the discrete logarithm problem on elliptic curves over
finite fields. The contribution in this part is a polynomial time reduction of the ECDLP on certain elliptic curves to the DLP in finite fields, which is established in Chapter 6. We now summarize the contents and the results.

In Chapter 2, we give the mathematical background needed for the next two chapters. We review some basic facts about finite fields and their arithmetic. All the results and the algorithms stated there are well known. A standard reference in the literature is the book by Lidl and Niederreiter [36]. We also give a brief introduction to the complex-analytic techniques that are used in later chapters for the analysis of our algorithm. They involve power series and the asymptotic estimate of their coefficients using the Cauchy formula and the saddle-point method, a standard technique for estimating the resulting integrals.

In Chapter 3, we prove some density theorems that are immediately applicable to the analysis of the algorithm of Chapter 4. We estimate the number of monic polynomials of degree \( n \) over a finite field \( \mathbb{F}_q \) that have only irreducible factors of degrees in a prescribed interval \( (m_2, m_1] \). Let \( N_q(n, m_1, m_2) \) be this number. If \( m_1 \) is a fixed constant, then the problem is of little interest for us, and can be solved using singularity analysis (see [49]). We deal with the case that \( m_1 \) tends to infinity as a function of \( n \). For \( m_2 \), we have two possibilities. If \( m_2 \) is a fixed constant and \( \sqrt{n} \log n \ll m_1 \ll n \), we obtain (in Theorem 3.3.1) a generalization of the theorems of Odlyzko [48] and Lovorn [37]. Furthermore, our estimate is in terms of the well known Dickman function, which establishes the interesting fact that the same function that governs the behavior of smooth integers and smooth polynomials also controls the behavior of a more general class of polynomials.

In order to increase the range of \( m_1 \), we weaken the result, to obtain in Theorem 3.3.2 an analog to the theorem of Canfield, Erdös and Pomerance [10]. Next, we study the case that \( m_2 \) also tends to infinity as a function of \( n \). In Theorem 3.4.3, we obtain asymptotic formulas for \( N_q(n, m_1, m_2) \) as \( \sqrt{n} \log n \ll m_1 \ll n \), and again a weaker result for a much greater range for \( m_1 \). The weak results are precisely those needed for the analysis of the algorithm, and are of the form

\[
N_q(n, m_1, m_2) = q^n \exp \left( -(1 + o(1)) \frac{n}{m_1} \log \frac{n}{m_1} \right),
\]

for \( \log n \ll m_1 \ll n \), and \( 0 \leq m_2 \leq cm_1 \) for any constant \( c < 1 \).

In Chapter 4, we present a generalized version of the index calculus method. Our intention is to investigate how the algorithm performs under different factor bases. The factor base is one of the most significant parameters of the method, and yet it has been taken as a fact, that the only good choice is the smooth factor base — i.e., the one that contains all irreducibles of degree up to some bound. This assumption has neither been proven nor challenged until now. We use the density theorems of Chapter 3, to analyze
a version of the index calculus that uses a non-smooth factor base, and conclude that theoretically it performs exactly the same as the version with the smooth factor base. Thus, the smooth factor base seems to offer no advantage. In practice, the choice of the factor base does make a difference. Some experiments that are reported at the end of the chapter indicate that the algorithm runs faster if the smooth factor base is used. Still, a non-smooth factor base might be the one of choice, if the space is a limited resource: the way to reduce the space requirements is to use a non-smooth factor base.

Chapter 5 presents an introduction to elliptic curves over finite fields. We define elliptic curves and state the main concepts such as the function field, the divisor group, the divisor class group, and the group law for an arbitrary base field \( K \). Then, we focus on the case \( K = \mathbb{F}_q \), and discuss some special properties that arise in this case. Our intention is to give a concrete presentation of the notions that are relevant to the arithmetic of elliptic curves, which will be used throughout Chapter 6.

Chapter 6 contains our main results on the discrete logarithm problem on elliptic curves. We start by reviewing the reduction of Menezes, Okamoto, and Vanstone [42] (MOV), that is based on the Weil pairing, and works for the class of supersingular curves, and discuss its limitations. Then, we briefly describe the reduction of Frey, Müller, and Rück [19] (FR). This reduction draws on a deeper mathematical result, namely the Tate-Lichtenbaum pairing. Our results are based on a generalization of the Weil pairing as it appears in [64, p.107]. It can, therefore, be viewed as a generalization of the MOV reduction. The generalization yields a pairing that is the inverse of the Tate-Lichtenbaum pairing, and therefore works for exactly the same cases that the FR reduction does, while being conceptually simpler.

Let \( P \in E(\mathbb{F}_q) \) be a point of prime order \( r \), and \( Q \in (P) \). Typically, the point \( P \) is chosen so that \( r \) is close to the order of \( E(\mathbb{F}_q) \) itself. In Section 6.3, we construct the generalized Weil pairing, and prove some properties that are crucial later. In Section 6.4, we focus on curves with \( a_q = 2 \), and we choose the parameters associated to the previous construction to obtain an isomorphism

\[
(P) \xrightarrow{\sim} \mu_r,
\]

where \( \mu_r \) is the group of \( r \)-th roots of unity in \( \mathbb{F}_q \). In Section 6.4.3, we give a polynomial time algorithm for computing the pairing. We conclude the chapter with a generalization of our approach that makes it work under the weaker condition \( r \mid q - 1 \). This condition is clearly necessary for any isomorphism from \( (P) \) to a subgroup of \( \mathbb{F}_q^\times \). Our construction shows it is also sufficient. The proof for the efficiency of our algorithm (for computing the pairing) is given in the appendix.
Finally, in Chapter 7, we discuss some directions for further research. They include possible further attacks on the elliptic curve discrete logarithm problem, as well as problems related to the DLP in groups not considered in this thesis, such as the ideal class group of number fields.
Chapter 2

Preliminary topics

In this chapter, we present the algebraic and analytic background that is needed for the first part of the thesis. The algebraic part focuses on the notions of groups, ring, and fields, and in particular finite fields, since this is the algebraic setting where everything takes place. In the analytic part, we present a counting technique and discuss the tools that are used. They include convergent power series, the Cauchy coefficient formula, and the saddle-point method for estimating integrals.

2.1 Algebraic preliminaries

The basic algebraic structure relevant to the discrete logarithm problem is that of a group. A subset of a group G that is also a group is called a subgroup of G. Here are some examples of groups.

Examples. The integers with the usual addition form a group. The identity is 0, and the inverse of $n \in \mathbb{Z}$ is $-n \in \mathbb{Z}$.

An example of a finite group is the set of non-zero integers modulo a prime $p$. The identity is 1, and the inverse of $a \in (\mathbb{Z}/p\mathbb{Z})^*$ is $a^{p-2}$ by Fermat’s little theorem.

A group $G$ is called cyclic if there is an element $g \in G$, such that all the elements of $G$ are of the form $g^n$ for some $n \in \mathbb{Z}$. In this case, $g$ is called a generator of $G$, and the group it generates is denoted by $\langle g \rangle$.

We focus now on finite groups. The number of elements of the group, $\#G$, is called the order of $G$. Then it follows that for every element $a \in G$ there exists a non-negative integer $n$ such that $a^n = 1$. The least such integer $n$ is called the order of $a$. We note that the order of $a$ is equal to the order of the subgroup $\langle a \rangle$ that $a$ generates. A fundamental
result about finite groups is Lagrange's theorem.

**Theorem 2.1.1** If $H$ is a subgroup of a finite group $G$, then $\#H$ divides $\#G$. Furthermore, if $n = \#G$, then for every element $a \in G$, $a^n = 1$.

An operation that is often performed when working with finite groups is exponentiation of elements. Let $a \in G$, and $k \geq 1$. We want to compute $a^k$. Lagrange's theorem implies that $a^k = a^{k \mod n}$, so we may assume that $k \leq n$. The idea is to use repeated squaring to compute $a, a^2, \ldots, a^{2^d}$ for $d = \lfloor \log_2 k \rfloor$, and then multiply the appropriate powers to obtain the result. More concretely, let $k = e_0 + e_1 2 + e_2 2^2 + \cdots + e_d 2^d$ be the number $k$ written in base 2. Then

$$a^k = a^{\sum_{i=0}^d e_i 2^i} = \prod_{i=0}^d \left( a^{2^i} \right)^{e_i}.$$  

We compute the powers $a^{2^i}$, $i = 0, \ldots, d$, by repeated squaring, which takes $O(d) = O(\log k)$ operations in $G$. Then we multiply the powers that correspond to the bits of the binary representation of $k$ that are equal to 1. This takes at most another $O(\log k)$ operations in $G$. Thus in total, any power $a^k$ can be computed using $O(\log n)$ operations in $G$, where $n = \#G$.

The next algebraic structure that we consider is the ring. Examples of commutative rings are the set of integers together with the usual addition and multiplication, and the set of integers modulo an integer $m$, $\mathbb{Z}/m\mathbb{Z}$.

Let $R$ be a commutative ring with multiplicative identity, and denote by $R^*$ the set $R - \{0\}$. If it happens that $(R^*, \cdot)$ is also a group, then $R$ is said to be a field. Thus, a field is a commutative ring, where every non-zero element has a multiplicative inverse. Common examples of fields are $\mathbb{Z}, \mathbb{Q}, \mathbb{R}$, and $\mathbb{C}$. They are all infinite fields.

An example of a field with finitely many elements is $\mathbb{F}_p = \mathbb{Z}/p\mathbb{Z}$, with $p$ a prime. In fact, for every prime $p$, $\mathbb{Z}/p\mathbb{Z}$ is a finite field. It can be proved that if $p$ is a prime, and $1 \leq n \in \mathbb{Z}$, then there exists a unique, up to isomorphism, finite field with $p^n$ elements. It is denoted by $\mathbb{F}_{p^n}$. We stress that $\mathbb{F}_{p^n}$ is not $\mathbb{Z}/p^n\mathbb{Z}$ for $n > 1$, as the latter is not a field. Furthermore, these are all the finite fields. The number $p$ is called the characteristic of $\mathbb{F}_{p^n}$.

### 2.1.1 Finite fields

We review now some basic facts about finite fields. For a thorough treatment of finite fields and their applications, we point to the book by Lidl and Niederreiter [36]. For the
remainder of the section, $q$ will denote a prime power. Let $\mathbb{F}_q$ be a finite field. Then $\mathbb{F}_q$ is by definition a multiplicative group. In fact it is a cyclic group of order $q - 1$. Thus, every element $a \in \mathbb{F}_q^*$ is of the form $g^n$ for some generator $g$ and a non-negative integer $n$.

Prime finite fields, i.e., fields of the form $\mathbb{F}_p$ with $p$ prime, are isomorphic to $\mathbb{Z}/p\mathbb{Z}$, so we have a representation for doing arithmetic. We develop now a representation for extension fields, i.e., fields with $p^n$ elements, $n \geq 2$, that will allow us to do arithmetic in those fields as well.

We consider the polynomial ring $\mathbb{F}_p[X]$. There exist irreducible polynomials in $\mathbb{F}_p[X]$ of any given degree $n$. Furthermore, it is a unique factorization domain — that is, every polynomial in $\mathbb{F}_p[X]$ can be written as the product of irreducible polynomials in essentially one way. If $f \in \mathbb{F}_p[X]$ is an irreducible polynomial of degree $n$, then $(f)$ is a maximal ideal of $\mathbb{F}_p[X]$, and therefore, $\mathbb{F}_p[X]/(f)$ is a field. It is not hard to see that it contains $p^n$ elements. It is therefore an isomorphic copy of $\mathbb{F}_{p^n}$. We will use the isomorphism

$$\mathbb{F}_q \cong \frac{\mathbb{F}_p[X]}{(f)},$$

for $q = p^n$ throughout this thesis. The elements of $\mathbb{F}_q$ are represented by polynomials of degree at most $n - 1$, and the arithmetic is done modulo the irreducible $f$. The following well-known estimate for the number of monic irreducible polynomials over $\mathbb{F}_q$ of given degree will be useful.

**Theorem 2.1.2** Let $\mathbb{F}_q$ be a finite field with $q$ elements. The number $I_n$ of monic irreducible polynomials over $\mathbb{F}_q$ of degree $n$ satisfies

$$I_n = \frac{q^n}{n} + O(q^{n/2}).$$

In addition to the above isomorphism, the index calculus method — which we study in Chapter 4 — uses a subroutine for factoring polynomials over $\mathbb{F}_p$. The problem of polynomial factorization has been studied since Gauss. As far as deterministic algorithms are concerned, the problem seems to be hard, since no polynomial time algorithms are known. We do, however, know randomized algorithms that factor a polynomial of degree $n$ over $\mathbb{F}_p$ in using a number of operations in $\mathbb{F}_p$ which is polynomial in $n$ and $\log p$. For results on factorization of polynomials over finite fields we refer to [49], and to [26] for an extensive survey on the subject. Finally, for a thorough treatment of finite fields and their applications, we point to the book by Lidl and Niederreiter [36].
2.2 Analytic and combinatorial techniques

The results needed for the analysis of our generalization of the index calculus involve estimates of the number $N_q(n, m_1, m_2)$ of monic polynomials over $\mathbb{F}_q$ of degree $n$, with all their irreducible factors having degree in the interval $(m_2, m_1]$. The counting techniques that we use are analytic. In this section, we review some of the basic notions related to those techniques.

2.2.1 Generating functions

In order to estimate $N_q(n, m_1, m_2)$, we first encode them as coefficients of formal power series. Let $R$ be a commutative ring with 1. A formal power series is an expression of the form

$$f(T) = \sum_{n=0}^{\infty} a_n T^n = a_0 + a_1 T + a_2 T^2 + \ldots,$$

with all coefficient $a_n \in R$. Let $f(T) = \sum a_n T^n$ and $g(T) = \sum b_n T^n$ be two power series over $R$. We define their sum to be

$$(f + g)(T) = \sum_{n=0}^{\infty} (a_n + b_n) T^n,$$

and their product

$$(fg)(T) = \sum_{n=0}^{\infty} \left( \sum_{k=0}^{n} a_k b_{n-k} \right) T^n.$$

Those operations make the set of formal power series over $R$ a ring. Furthermore, we have the following theorem about the units (invertible elements) in this ring.

**Theorem 2.2.1** A formal power series $f(T)$ has an inverse if and only if it has non-zero constant term.

For example, we see that the power series $1 + T + T^2 + T^3 + \ldots$ has an inverse, namely, $1 - T$. Indeed, one can verify that

$$(1 + T + T^2 + T^3 + \ldots) \cdot (1 - T) = 1.$$

We can also compose power series, as we do with polynomials. Let $f(T) = a_0 + a_1 T + a_2 T^2 + \ldots$ be a power series and $g(T) = b_1 T + b_2 T^2 + \ldots$ be power series with zero constant term. Then, the composite $f \circ g(T) = f(g(T)) = a_0 + a_1 g + a_2 g^2 + \ldots$ is well defined.
Let \((a_n)_{n \in \mathbb{N}}\) be a sequence of complex numbers. Then, the power series

\[
\sum_{n=0}^{\infty} a_n T^n
\]

is called the generating function of the sequence \((a_n)_{n \in \mathbb{N}}\).

We proceed now, to demonstrate how to obtain the generating function for the numbers \(N_q(n, m_1, m_2)\). It is the formal power series

\[
\sum_{n=0}^{\infty} N_q(n, m_1, m_2) T^n.
\]

To avoid confusion, we adopt the following notation: The monic polynomials in \(\mathbb{F}_q[X]\) are denoted by boldface letters, like \(f, g\), etc. The letter \(p\) will always stand for a monic irreducible polynomial over \(\mathbb{F}_q\). Then we compute

\[
\sum_{n=0}^{\infty} N_q(n, m_1, m_2) T^n = \sum_{n=0}^{\infty} \sum_{\deg(f) = n, p \mid f \Rightarrow m_2 < \deg(p) \leq m_1} T^{\deg(f)} \tag{2.1}
\]

On the other hand, we have

\[
\prod_{m_2 < \deg(p) \leq m_1} \frac{1}{1 - T^{\deg(p)}} = \prod_{m_2 < \deg(p) \leq m_1} (1 + T^{\deg(p)} + T^{2 \deg(p)} + \cdots) = 1 + \sum_{p \mid f \Rightarrow m_2 < \deg(p) \leq m_1} T^{\deg(f)}.
\]

The first equality is an identity of formal power series, and the second holds due to the unique factorization of polynomials over finite fields. If we group now the terms according to the degree, we get

\[
\prod_{m_2 < \deg(p) \leq m_1} \frac{1}{1 - T^{\deg(p)}} = 1 + \sum_{n=0}^{\infty} \sum_{\deg(f) = n, p \mid f \Rightarrow m_2 < \deg(p) \leq m_1} T^{\deg(f)} \tag{2.2}
\]

From Equations (2.1) and (2.2) we get

\[
1 + \sum_{n=0}^{\infty} N_q(n, m_1, m_2) T^n = \prod_{m_2 < \deg(p) \leq m_1} \frac{1}{1 - T^{\deg(p)}} = \prod_{m_2 < k \leq m_1} (1 - T^k)^{-f_k},
\]

where \(f_k\) is the number of monic irreducible polynomials over \(\mathbb{F}_q\) of degree \(k\).
2.2.2 Convergent power series

For the rest of the chapter, we focus on power series over \( \mathbb{C} \). The properties of formal power series are usually not enough to give estimates for the coefficients. In the case that a formal power series converges, it defines a function on \( \mathbb{C} \), and we can use methods of complex analysis to study it. We start with the following basic theorem.

**Theorem 2.2.2** Let \( \sum a_n z^n \) be a power series. If it does not converge absolutely for all \( z \in \mathbb{C} \), then there exists a number \( r \) such that the series converges absolutely for \( |z| < r \) and does not converge for \( |z| > r \). The number \( r \) is called the radius of convergence of the power series.

Thus, inside the circle of convergence, power series define analytic functions (differentiable). Power series can be differentiated term by term in any region and integrated along any path that lies entirely inside the circle of convergence.

We consider now a function in one complex variable, \( f(z) \), that is analytic inside and on a simple closed curve \( C \). Let \( a \) be any point inside \( C \). Then Cauchy's formula gives the values of the derivatives \( f^{(n)}(a) \) as the integral along the closed curve \( C \) (traveled counter-clockwise),

\[
f^{(n)}(a) = \frac{n!}{2\pi i} \int_C \frac{f(z)}{(z-a)^{n+1}} dz.
\]

In our case, the generating function

\[
P(z) = 1 + \sum_{n=0}^{\infty} N_q(n, m_1, m_2) z^n = \prod_{m_2 < k \leq m_1} (1 - z^k)^{-l_k}.
\]

Thus, Cauchy’s formula, applied with \( a = 0 \), states that

\[
P^{(n)}(0) = \frac{n!}{2\pi i} \int_C \frac{P(z)}{z^{n+1}} dz.
\]

We note now that the \( n \)-th coefficient of \( P(z) \), denoted by \( [z^n] P(z) \), can be recovered as

\[
N_q(n, m_1, m_2) = [z^n] P(z) = \frac{1}{n!} P^{(n)}(0) = \frac{1}{2\pi i} \int_C \frac{P(z)}{z^{n+1}} dz.
\]

It turns out, that it is more convenient to consider the function

\[
P\left(\frac{z}{q}\right) = \sum_{n=0}^{\infty} N_q(n, m_1, m_2) \frac{z^n}{q^n}.
\]

If \( N_q(n, m) \) denotes the number of monic polynomials over \( \mathbb{F}_q \) of degree \( n \) with all irreducible factors having degree at most \( m \), then clearly

\[
N_q(n, m_1, m_2) \leq N_q(n, m_1).
\]
Furthermore, it is known (see [68]) that for \( 2(\log n)^2 \leq m \leq n \),

\[
N_q(n, m) = q^n e^{-\frac{n}{m} \log \left( \frac{n}{m} \right)} e^{O\left( \frac{n}{m} \log n \right)}.
\]

Thus, the power series \( P(z/q) \) converges absolutely for

\[
|z| < \lim_{n \to \infty} \frac{q}{N_q(n, m)^{1/n}} = q.
\]

The Cauchy coefficient formula gives

\[
N_q(n, m_1, m_2) = q^n [z^n] P \left( \frac{z}{q} \right) = \frac{1}{2\pi i} \int_{C'} \frac{P(z/q)}{z^{n+1}} dz.
\]

Our contour of integration \( C' \) will be roughly \( |z| = e^{1/m_1 \log n} \), which as we showed, is within the circle of convergence of \( P(z/q) \). We note that the proofs of Theorems 3.4.3 and 3.4.5 use analytic properties of \( P(z/q) \), and therefore the condition \( 2(\log n)^2 \leq m_1 \leq n \), that we used to show that the contour \( C' \) lies inside the circle of convergence of the power series will have to be satisfied.

It remains the problem of estimating the integral in Equation (2.3). For that we use the saddle point method. For a general description of the method we refer to the book by DeBruijn [16]. We find it easiest to explain the method, as it applies to our case, during the course of the proof, in Chapter 3.
Chapter 3

Density theorems

3.1 Motivation and results

This chapter is devoted to the study of certain factorization patterns of polynomials over finite fields. Let \( \mathbb{F}_q[X] \) be the ring of polynomials in one variable with coefficients in \( \mathbb{F}_q \). It is well-known that \( \mathbb{F}_q[X] \) is a unique factorization domain; that is, every nonzero polynomial \( f \in \mathbb{F}_q[X] \) can be written in a unique way (up to term rearrangement, and multiplication by elements in \( \mathbb{F}_q \)) as the product of irreducible polynomials over \( \mathbb{F}_q \), say

\[
f = \prod_p p^{v_p(f)}.
\]

If we restrict our attention to monic polynomials, then we do not need to worry about irreducibles that differ by a multiplicative constant. We are interested in estimating the number \( N_q(n, m_1, m_2) \) of monic polynomials of degree \( n \), all of whose irreducible factors are of degree greater than \( m_2 \) and at most \( m_1 \). This condition can be written as

\[
p | f \implies m_2 < \deg(p) \leq m_1.
\]

The problem is of interest from a purely mathematical point of view, as well as for its application to the analysis of the index calculus method. More generally, if \( R \) is a set of structures that decompose in a unique way into "irreducible" structures, and we have a size function \( | \cdot | : R \rightarrow \mathbb{R} \), then the same question can be formulated in \( R \). The algebraic domains that have mainly been considered are the ring of integers \( \mathbb{Z} \) (integers decompose into prime factors), the polynomial ring \( \mathbb{F}_q[X] \) (polynomials decompose into irreducible factors), and the set of integral ideals of the ring of integers in a number field (integral ideals decompose into prime ideals). A general treatment on the subject appears in the book by Knopfmacher [32].
In the context of integers, first order asymptotic results for \( m_2 = 0 \), i.e., for the so-called smooth integers, have been obtained by De Bruijn [15], Hildebrand and Tenenbaum [31], and others. Canfield, Erdős, and Pomerance [10] obtained weaker results that hold for a wider range of values for \( m_1 \) (than De Bruijn's estimates). Friedlander [21] considered the problem for varying \( m_2 \).

In the context of polynomials over \( \mathbb{F}_q \), the problem has been studied only in the case \( m_2 = 0 \), i.e., only for smooth polynomials. Both first order asymptotic results as well as weaker approximations have been obtained by Odlyzko [48], Lovorn [37], Lovorn Bender and Pomerance [38], Soundararajan [68], and others.

Finally, smooth ideals have attracted far less attention. The sharpest results are obtained by Buchmann and Hollinger [9]. They are upper bounds in the case \( m_2 = 0 \).

Our study focuses on polynomials over \( \mathbb{F}_q \). The estimates for \( N_q(n, m_1, m_2) \) that we obtain hold for \( q \) fixed as \( n \to \infty \). The notation in this chapter is as follows. We write \( f(n) \ll g(n) \) for two functions \( f, g \) if \( \lim_{n \to \infty} f(n)/g(n) = 0 \). First we consider the case that \( m_2 \) is any fixed constant (independent of \( n \)), and prove a first order approximation for \( \sqrt{n} \log n \ll m_1 \ll n \) (Theorem 3.3.1). A weaker estimate that holds for \( \log n \ll m_1 \ll n \) (Theorem 3.3.2) follows easily. Then we allow \( m_2 \) to tend to infinity as a function of \( n \), and we obtain again first order approximations for \( \sqrt{n} \log n \ll m_1 \ll n \) (Theorem 3.4.3 and Corollary 3.4.4), and weaker results for \( 2(\log n)^2 \le m_1 \ll n \) (Theorem 3.4.5). We note that the weak estimates are enough for the analysis of our version of the index calculus, that is presented in Chapter 4. The results of this chapter can be found in [24]

### 3.2 A technical lemma

In this section we prove a technical lemma that we need for the proof of Theorem 3.3.1 of Section 3.4. It will be crucial to estimate the part of the logarithm series between \( m_1 \) and \( m_2 \)

\[
r_{m_1,m_2}(z) = \sum_{k=m_2+1}^{m_1} \frac{z^k}{k}.
\]

The following lemma, extension of that in [50], provides an estimate for \( r_{m_1,m_2}(z) \) in terms of the exponential integral

\[
E(a) = \int_a^\infty \frac{e^{-s}}{s} ds.
\]

**Lemma 3.2.1** Let \( n, m_1, m_2 \in \mathbb{N} \), and let \( u_1 = n/m_1 \), \( u_2 = n/m_2 \). Assume that all the above quantities tend to infinity. Let \( h = -\xi + i\tau \), with \( \xi > 0 \) and \( \xi/n \to 0 \).
If \(|\tau| \leq u_2\), then
\[
 r_{m_1,m_2}(e^{-h/n}) = E(h/u_2) - E(h/u_1) + O\left(\frac{\xi + \tau}{n} e^{\xi/u_1}\right). \tag{3.1}
\]

For any value of \(\tau\),
\[
 r_{m_1,m_2}(e^{-h/n}) = O(E(h/u_2) - E(h/u_1)). \tag{3.2}
\]

**Proof.** By definition of \(r_{m_1,m_2}(z)\), we have
\[
 r_{m_1,m_2}(e^{-h/n}) = \sum_{k=m_2+1}^{m_1} e^{-kh/n} = \sum_{k=m_2+1}^{m_1} \int_{h/n}^{\infty} e^{-ky} dy
\]
\[
 = \int_{h/n}^{\infty} \left(\sum_{k=m_2+1}^{m_1} e^{-ky}\right) dy = \int_{h/n}^{\infty} \frac{e^{-m_2y}}{e^y - 1} dy - \int_{h/n}^{\infty} \frac{e^{-m_1y}}{e^y - 1} dy
\]
\[
 = \int_{h/u_2}^{\infty} e^{-s} \frac{1/m_2}{e^{s/m_2} - 1} ds - \int_{h/u_1}^{\infty} e^{-s} \frac{1/m_1}{e^{s/m_1} - 1} ds.
\]

Consider now the integral
\[
 \int_{h/u_2}^{\infty} e^{-s} \frac{1/m_2}{e^{s/m_2} - 1} ds = \int_{h/u_1}^{\infty} e^{-s} \frac{s/m_1}{e^{s/m_1} - 1} ds,
\]
where \(u = n/m\), and let \(\psi(z) = \frac{e^z - 1}{z}\), which is analytic for \(|z| < 2\pi\). Then, the above integral can be written as
\[
 \int_{h/u}^{\infty} e^{-s} \frac{s}{e^{s/m} - 1} ds = \int_{h/u}^{\infty} e^{-s} \frac{s/m}{e^{s/m} - 1} ds.
\]

We recall that \(\Re(h) = -\xi < 0\). The term \(E(h/u)\) is the main term in the approximation of Equation (3.1). To finish the proof of Equation (3.1) we need to bound the last two integrals. For the first integral in Equation (3.3), we observe that for \(s\) in the range \((h/u, -h/u)\), we have
\[
 \frac{|s|}{m} < \frac{|h|}{m u} \leq \frac{1}{m} \left(\frac{m \xi + |\tau|}{u}\right)
\]
\[
 \leq \frac{\xi}{n} + \frac{|\tau|}{n} \leq \frac{\xi}{n} + \frac{1}{m_2},
\]
where in the last step we used the assumption $|\tau| \leq u_2$ of the lemma. Since $\xi / n \to 0$, and $m_2 \to \infty$, we get $|h| / m \to 0$. Therefore, expanding $e^{s/m}$ we obtain

$$\psi(s/m) - 1 = \frac{s/m}{e^{s/m} - 1} - 1 = \frac{s/m}{s/m + O(s^2/m^2)} - 1 = \frac{1}{1 + O(s/m)} - 1 = O(s/m).$$

This implies

$$\left| \int_{-h/u}^{-h/u} \frac{e^{-s}}{s} \left( \psi \left( \frac{s}{m} \right) - 1 \right) \, ds \right| \leq \frac{e^{\xi/u} 2|h|}{|h/u|} O \left( \frac{h}{n} \right) = O \left( e^{\xi/u} \frac{\xi + \tau}{n} \right),$$

where the last equality holds, since $|h| = O(\xi + \tau)$. For the second integral in Equation (3.3), we note that in the range $(-h/u, \infty)$ the function $\psi$ is bounded (a proof of that is provided below for the second part of the lemma). Therefore,

$$\left| \int_{-h/u}^{\infty} \frac{e^{-s}}{s} \left( \psi \left( \frac{s}{m} \right) - 1 \right) \, ds \right| \leq O(1) E(-h/u).$$

Next we need to bound $E(-h/u)$. For that we establish the following bound, as the real part $\sigma$ of the argument is positive,

$$E(\sigma + i\tau) = \int_{\sigma + i\tau}^{\infty} \frac{e^{-s}}{s} \, ds = \int_{\sigma + i\tau}^{\sigma} \frac{e^{-s}}{s} \, ds + \int_{\sigma}^{\infty} \frac{e^{-s}}{s} \, ds. \tag{3.4}$$

The second integral is $O(e^{-\sigma}/\sigma)$. The first integral, after the substitution $s = \sigma + iy$, becomes

$$\int_{\sigma + i\tau}^{\sigma} \frac{e^{-s}}{s} \, ds = \int_{\tau}^{0} \frac{e^{-iy}}{\sigma + iy} \, dy.$$

One can check now that the second integral is $O(e^{-\sigma} \log \tau)$.

Returning to the lemma now, we conclude that the integral in the range $(-h/u, \infty)$ is $O(e^{-\xi/u} \log(\tau/u))$, which is absorbed by the error term induced by the integral in the range $(h/u, -h/u)$ (recall that $\xi > 0$). Equation (3.1) now follows by subtraction, considering $u = u_2$ and $u = u_1$, and noting that $u_1 < u_2$.

We concentrate in Equation (3.2). We observe that $\psi(z)$ is analytic for $|\Re(z)| < 2\pi$, and for the range of integration this condition holds for $s/m$. Furthermore, we have

1. for $s = o(m)$, $\psi(s/m) = 1 + o(1)$,

2. for $s = cm$, $\psi(s/m) = \psi(c) = \frac{e^{c/m}}{e^{c/m} - 1}$. 

3. for $s = \omega(m)$, $\psi(s/m) \to 0$,

which proves that $\psi(s/m) = O(1)$. Equation (3.2) now follows immediately from

$$r_{m_1,m_2}(e^{h/n}) = \int_{h/u_2}^{\infty} \frac{e^{-s}}{s} \psi\left(\frac{s}{m_2}\right) ds - \int_{h/u_1}^{\infty} \frac{e^{-s}}{s} \psi\left(\frac{s}{m_1}\right) ds.$$

The choice of the parameters of the above lemma, although somewhat artificial, is made to fit exactly the saddle point method that is extensively used in section 3.4.

### 3.3 Fixed lower bound

We turn now to the problem of counting the number of monic polynomials over a finite field that are free of irreducible factors of small and large degree. We start by fixing the notation. Let $q$ be a prime power, and $\mathbb{F}_q$ be the finite field with $q$ elements. We use boldface letters, e.g., $\mathbf{f}$, $\mathbf{d}$, to denote polynomials over $\mathbb{F}_q$. In particular, $\mathbf{p}$ always denotes an irreducible polynomial, and $\mathbf{1}$ denotes the unit of $\mathbb{F}_q$. We are interested in the number of monic polynomials of degree $n$ over $\mathbb{F}_q$ with all irreducible factors having degree greater than $m_2$ and less than or equal to $m_1$, which we denote by $N_q(n, m_1, m_2)$.

In general both $m_1$ and $m_2$ can (and will) be functions of $n$.

In this section we consider the case when the lower bound $m_2$ for the degree of the irreducible factors is a fixed constant. The special case when $m_2 = 0$ corresponds to the smooth polynomials [48]. The number of generalized smooth polynomials, i.e., when $m_2$ is fixed, under certain conditions on $m_1$, can be expressed in terms of the Dickman function which also governs the behavior of smooth integers [15, 31]. The Dickman function is defined as the unique solution of the following difference-differential equation:

$$\begin{align*}
\rho(u) &= 1, \quad 0 \leq u \leq 1 \\
u \rho'(u) &= -\rho(u - 1), \quad u > 1 \\
\rho(u) \text{ is continuous.}
\end{align*}$$

The result in this section is expressed in terms of the M"obius function $\mu$ that is defined as follows:

$$\begin{align*}
\mu(1) &= 1 \\
\mu(p) &= -1 \\
\mu(p^k) &= 0, \quad \text{for } k \geq 2
\end{align*}$$
and it is extended multiplicatively to all polynomials, i.e.,

\[ \text{if } \gcd(f, g) = 1, \quad \mu(fg) = \mu(f)\mu(g). \]

The following theorem gives an asymptotic estimate for \( N_q(n, m_1, m_2) \). The estimate can be established using analytic methods. An argument similar to the one used for the proof of Theorem 3.4.3 would work. However, the fact that \( m_2 \) is constant allows for an elementary argument that we now give.

**Theorem 3.3.1** The number \( N_q(n, m_1, m_2) \) of monic polynomials of degree \( n \) over \( \mathbb{F}_q \) with all irreducible factors with degree greater than \( m_2 \) and less than or equal to \( m_1 \), with \( m_2 \) fixed and \( \sqrt{n} \log n \ll m_1 \ll n \) satisfies

\[
N_q(n, m_1, m_2) \sim q^n \rho \left( \frac{n}{m_1} \right) \sum_{d \mid \gcd(f, P(m_2))} \mu(d) q^{\deg(d)},
\]

where \( \rho \) is the Dickman function, \( \mu \) is the Möbius function, and \( P(m_2) = \prod_{p \mid \deg(p) \leq m_2} P \), where the product is over irreducibles \( p \).

**Proof.** Let \( N_q(n, m) \) denote the number of \( m \)-smooth polynomials of degree \( n \). We have

\[
N_q(n, m_1, m_2) = \sum_{\substack{r \mid f \mid m_2 < \deg(p) \leq m_1 \\text{ and} \\deg(f) = n}} 1 = \sum_{d \mid \gcd(f, P(m_2))} \mu(d) \sum_{\substack{r, f \mid m_2 < \deg(p) \leq m_1 \\text{ and} \\deg(f) = n}} 1 = \sum_{d \mid \gcd(f, P(m_2))} \mu(d) N_q(n - \deg(d), m_1). \quad (3.6)
\]

It is well known (see [39]) that in the range \( \sqrt{n} \log n \ll m \leq n \), \( N_q(n, m) = (1 + o(1))q^n \rho(n/m) \). Using this and Equation (3.6) we get

\[
N_q(n, m_1, m_2) = \sum_{d \mid \gcd(f, P(m_2))} (1 + o(1)) \mu(d) q^n \rho \left( \frac{n - \deg(d)}{m_1} \right) q^{\deg(d)} \quad (3.7)
\]

\[
= q^n \rho \left( \frac{n}{m_1} \right) \sum_{d \mid P(m_2)} \frac{\mu(d)}{q^{\deg(d)}} + o \left( q^n \rho \left( \frac{n}{m_1} \right) \sum_{d \mid P(m_2)} \frac{\mu(d)}{q^{\deg(d)}} \right),
\]

because \( \rho((n - \deg(d))/m_1) \sim \rho(n/m_1) \), since \( \deg(d) \) is a constant. \( \Box \)

In the spirit of Canfield, Erdös and Pomerance [10], we can obtain a slightly weaker result that holds for a much larger range of values of \( m_1 \). Indeed, replacing the assumption \( \sqrt{n} \log n \ll m_1 \) by \( \log n \ll m_1 \), the following estimate is well-known

\[
\rho(u) = e^{-(1+o(1))u \log u}, \quad u = \frac{n}{m}.
\]
Using this, Theorem 3.3.1, and observing that the sum \( \sum_{d \in \mathcal{P}(m_2)} \frac{\mu(d)}{d^{n+1}} \) is a constant (since \( m_2 \) is constant), the following theorem is immediate.

**Theorem 3.3.2** The number \( N_q(n, m_1, m_2) \) of monic polynomials of degree \( n \) over \( \mathbb{F}_q \) with all irreducible factors with degree between \( m_2 \) and \( m_1 \), with \( m_2 \) fixed and \( \log n \ll m_1 \ll n \), satisfies
\[
N_q(n, m_1, m_2) = q^n e^{-(1+o(1))u_1 \log u_1},
\]
where \( u_1 = n/m_1 \).

### 3.4 Lower bound tending to infinity

We turn now to the case when both bounds \( m_1 \) and \( m_2 \) tend to infinity, with \( m_1 = o(n) \). The proof of the main theorem of this section is based on analytic techniques. An elementary argument such as the one given in the previous section could be extended to work for \( m_2 \) tending to infinity very slowly. Technically, the point where the elementary argument breaks down is in Equation (3.7). The degree of the polynomial \( d \) is not constant any more, and therefore one does not have the approximation \( \rho((n - \deg(d))/m_1) \sim \rho(n/m_1) \).

Analytic arguments make possible an estimate that holds for a wider range for \( m_2 \). The main idea of the proof is to find the generating function \( P_{m_1, m_2}(z) \) of the numbers \( N_q(n, m_1, m_2) \), and then estimate the coefficients. We use the method outlined in Section 2.2.2 to express \( P_{m_1, m_2}(z) \) in terms of the sum \( r_{m_1, m_2}(z) = \sum_{k=m_2+1}^{m_1} \frac{z^k}{k} \). Then we use Lemma 3.2.1 to express it in terms of the exponential integral. The coefficients of the generating function are given by Cauchy integral, which is estimated via the saddle point method. The application of the saddle point method is the main and most involved part of the proof.

Let \( z \) be a formal variable. Let \( I_k \) denote the number of monic irreducible polynomials over \( \mathbb{F}_q \) of degree \( k \). We recall from Chapter 2 that the generating function is
\[
P_{m_1, m_2}(z) = \prod_{k=m_2+1}^{m_1} (1 - z^k)^{-I_k} = \exp \left( - \sum_{k=m_2+1}^{m_1} I_k \log(1 - z^k) \right)
= \exp \left( \sum_{j=1}^{\infty} \frac{1}{j} \sum_{k=m_2+1}^{m_1} I_k z^{jk} \right) = \exp \left( r_{m_1, m_2}^{[1]}(z) + \frac{r_{m_1, m_2}^{[2]}(z)}{2} + \cdots \right)
\]
where \( r_{m_1, m_2}^{[j]}(z) = \sum_{k=m_2+1}^{m_1} I_k z^{jk} \), \( j \geq 1 \). Using now the estimate \( k I_k = q^k + O(q^{k/2}) \), we obtain for \( |z| < q^{1/10} \)
\[
r_{m_1, m_2}^{[1]} \left( \frac{z}{q} \right) = \sum_{k=m_2+1}^{m_1} \frac{z^k}{k} + O(q^{-2m_2/5}),
\]
and

\[ r_{m_1,m_2}^{[1]} \left( \frac{z}{q} \right) = O \left( q^{(-9j/10+1)m_2} \right) = O \left( q^{-4m_2/5} \right), \quad j \geq 2. \]

The Cauchy coefficient formula now gives

\[
N_q(n, m_1, m_2) = \left[ z^n \right] P_{m_1, m_2}(z) = q^n \left[ z^n \right] P_{m_1, m_2} \left( \frac{z}{q} \right) = \frac{q^n}{2\pi i} \oint_C P_{m_1, m_2} \left( \frac{z}{q} \right) \frac{dz}{z^{n+1}}
\]

where the contour \( C \) is chosen to be \( z = e^{-\alpha/n+i\theta}, -\pi \leq \theta \leq \pi, \) and \( \alpha \) is a parameter to be chosen later. The idea for this substitution first appears in the thesis of Gourdon [28]. The change of variable \( z = e^{-h/n} \) implies \( h = \alpha - n\theta, \) and the limits of integration now are \((\alpha + in\pi, \alpha - in\pi).\) Therefore,

\[
N_q(n, m_1, m_2) = \frac{q^n}{2\pi i} \int_{\alpha + in\pi}^{\alpha - in\pi} P_{m_1, m_2} \left( \frac{e^{-h/n}}{q} \right) \frac{dh}{e^{-h}}
\]

where we used that \( r_{m_1,m_2}^{[1]}(z/q) = r_{m_1,m_2}(z) + o(1). \)

We focus on the above integral. Let us denote

\[
J(n, m_1, m_2) = \frac{1}{2\pi i} \int_{\alpha - in\pi}^{\alpha + in\pi} e^{r_{m_1,m_2}(e^{-h/n})} \frac{dh}{n}
\]

We expect that the main contribution to the integral comes from the neighborhood of the real axis. If the saddle point method is to work, then we hope to approximate the term \( r_{m_1,m_2}(e^{-h/n}) \) close to the real axis by \( E(h/u_2) - E(h/u_1) + O(\frac{x^\alpha}{n} e^{-\alpha/u_1}), \) according to Lemma 3.2.1, provided of course that the value of \( \alpha \) satisfies the conditions of the lemma. Let

\[
f(h) = E(h/u_2) - E(h/u_1) + h.
\]

The value of \( \alpha \) is determined as the real solution of the equation \( f'(h) = 0, \) that is,

\[
f'(h) = 1 - \frac{e^{-h/u_2}}{h} + \frac{e^{-h/u_1}}{h} = 0.
\]

The following lemma will be crucial for the application of the saddle point method.

**Lemma 3.4.1** The equation

\[
\frac{e^{-h/u_2}}{h} - \frac{e^{-h/u_1}}{h} = 1 \quad (3.10)
\]
has a negative real solution, \(-\xi\), such that \(\xi \sim u_1 \log(u_1 \log u_1)\). More precisely,

\[ u_1 \log(u_1 \log u_1) < \xi < u_1 \log(u_1 \log^2 u_1), \]

where \(u_1 = n/m_1, u_2 = n/m_2, m_2 \leq cm_1\) for any constant \(c < 1\), and \(u_1 \to \infty\).

**Proof.** Consider the function \(f'(h)\) at the points \(h_1 = -u_1 \log(u_1 \log u_1)\), and \(h_2 = -u_1 \log(u_1 \log^2 u_1)\). One can easily check that \(f'(h_1) > 0\), and \(f'(h_2) < 0\), and since \(f'(h)\) is continuous, it follows that it has a zero in \((h_2, h_1)\). \(\square\)

To estimate \(J(n, m_1, m_2)\) we choose \(\alpha = -\xi\). We break \(J(n, m_1, m_2)\) as the sum of two integrals along the vertical line with \(\Im(h) = -\xi\), as shown in Figure 3.1

\[ J(n, m_1, m_2) = \frac{1}{2n\pi i} \int_{-\xi-i\delta}^{-\xi+i\delta} e^{r_{m_1,m_2}(e^{-h/n})} dh + \frac{1}{2n\pi i} \int_{\delta < |\Im(h)| < \pi} e^{r_{m_1,m_2}(e^{-h/n})} dh. \]

Nearly all the contribution to the integral will come from the neighborhood of the point \(-\xi\), with the contribution of the tail being negligible. To show that, we first estimate

\[ \frac{1}{2n\pi i} \int_{-\xi-i\delta}^{-\xi+i\delta} e^{r_{m_1,m_2}(e^{-h/n})} \frac{eh}{n} dh, \]

for a suitable \(\delta\). One can easily check now that the conditions of Equation (3.1) in Lemma 3.2.1 are satisfied — in fact the conditions were chosen to fit the proof. Substituting the estimate \(f(h)\) for \(r_{m_1,m_2}(e^{-h/n}) + h\), we obtain

\[ \frac{1 + o(1)}{2n\pi i} \int_{-\xi-i\delta}^{-\xi+i\delta} \exp(f(h)) dh. \]

We need to estimate

\[ J_\delta(n, m_1, m_2) = \frac{1}{2n\pi i} \int_{-\xi-i\delta}^{-\xi+i\delta} \exp(f(h)) dh. \tag{3.11} \]

We need the first three derivatives at \(h = -\xi\). By the definition of \(\xi\), we have

\[ f'(-\xi) = 0. \]

For the second derivative we can write

\[ f''(-\xi) = \frac{1}{-\xi} \left( \frac{e^{\xi} u_2}{u_2} - \frac{e^{\xi} u_1}{u_1} + 1 \right) \]

\[ \approx \frac{1}{u_1 \log(u_1 \log u_1)} \left( \frac{(u_1 \log u_1)^{u_1/u_2}}{u_2} - \log u_1 + 1 \right) \]

\[ \approx \frac{-\log u_1}{u_1 \log(u_1 \log u_1)} \sim \frac{-1}{u_1}, \tag{3.12} \]
Figure 3.1: Decomposition of the integral
where we used the facts that $-\xi$ satisfies Equation (3.10), that $\xi \sim u_1 \log(u_1 \log u_1)$, and that $u_2 \geq cu_1$ for some $c < 1$. Finally, the third derivative can be computed and shown to be

$$f''''(-\xi) = O\left(\frac{e^{\xi/u_1}}{u_1^2 \xi}\right) = O\left(\frac{1}{u_1^2}\right).$$

It follows that

$$f(-\xi + it) = f(-\xi) - \frac{f''(-\xi)}{2} t^2 + O\left(f''''(-\xi)t^3\right).$$

Using Lemma 3.4.1 and for $\delta = u_1^{1/2} \log u_1$, we have $t^3 f''''(-\xi) = O\left(u_1^{-1/2}(\log u_1)^3\right)$.

Under the change of variable $h = -\xi + it$, we have

$$J_\delta(n, m_1, m_2) = \frac{1}{2\pi} \int_{-\delta}^{\delta} \exp\left(f(-\xi) - t^2 f''(-\xi)/2 + t^3 O\left(f'''(-\xi)\right)\right) dt$$

$$= \frac{\exp(f(-\xi))}{2n\pi} \int_{-\delta}^{\delta} \left(1 + t^3 O\left(f'''(-\xi)\right)\right) \exp\left(-t^2 f''(-\xi)/2\right) dt.$$  

The term containing $t^3$ is $o(1)$. Therefore, we obtain

$$J_\delta(n, m_1, m_2) = (1 + o(1)) \frac{\exp(f(-\xi))}{2n\pi} \int_{-\delta}^{\delta} \exp\left(-t^2 f''(-\xi)/2\right) dt$$

$$= (1 + o(1)) \frac{\exp(f(-\xi))}{n\sqrt{2\pi f''''(-\xi)}}. \tag{3.13}$$

For the tails now, we will use the following technical lemma.

**Lemma 3.4.2** For $h = -\xi + it$, $t \in \mathbb{R}$, and $g(h) = O(E(h/u_2) - E(h/u_1)) + h$, we have

$$\exp(g(h)) = \exp(-\xi + o(u_1) + it), \quad |t| \geq u_2; \tag{3.14}$$

$$\exp(\Re(f(h))) \leq \exp\left(f(-\xi) - \frac{Kt^2}{u_1}\right), \quad |t| \leq u_1; \tag{3.15}$$

$$\exp(\Re(f(h))) \leq \exp\left(f(-\xi) - \frac{u_1}{(\log u_1)^2 + \pi^2} + O(1)\right), \quad u_1 < |t| < u_2. \tag{3.16}$$

**Proof.** We note that $E(s) = o(e^{-s}/|\tau|)$, for $s = \sigma + i\tau$ (see [69], p. 373). Therefore, for $|t| \geq u_2$

$$\exp(g(h)) = \exp\left(O\left(e^{\xi/u_2}u_2/|t|\right) - O\left(e^{\xi/u_1}u_1/|t|\right) - \xi + it\right)$$

$$= \exp\left(-\xi + o(u_1) + it\right),$$

where we need to assume $u_1 \log u_1 = o(u_2)$, which is satisfied if $m_2 = o(m_1/\log n)$. 
The second equation is more involved. We start by checking that (see [1], 5.1.37)

\[
\begin{align*}
    f(-\xi) - \Re(f(h)) &= \int_0^1 \frac{e^{\xi y/u_1} (1 - \cos(ty/u_1))}{y} dy - \int_0^1 \frac{e^{\xi y/u_2} (1 - \cos(ty/u_2))}{y} dy \\
    &\geq \int_0^1 \frac{e^{\xi y/u_1} (\cos(ty/u_2) - \cos(ty/u_1))}{y} dy.
\end{align*}
\] (3.17)

For \(|t| \leq u_1\) and an appropriate constant \(K'\), we have

\[
\cos(ty/u_2) - \cos(ty/u_1) \geq \frac{K' t^2 y^2}{u_1^2}.
\] (3.18)

Using Equations (3.17) and (3.18), we have for a constant \(K\)

\[
\begin{align*}
    f(-\xi) - \Re(f(h)) &\geq \int_0^1 \frac{e^{\xi y/u_1} K' t^2 y}{u_1^2} dy = \frac{K' t^2}{u_1^2} \int_0^1 y e^{\xi y/u_1} dy \\
    &= \frac{K' t^2}{u_1^2} \frac{\xi}{\xi^2} \left( \frac{\xi - u_1}{u_1} u_1 + u_1^2 \right) \sim \frac{K' t^2}{u_1}.
\end{align*}
\]

The bound in Equation (3.16) can be shown in a similar way. We work in the range \(u_1 < |t| < u_2\). We observe that

\[
\begin{align*}
    f(-\xi) - \Re(f(h)) &= \int_0^1 \frac{e^{\xi y/u_1} (1 - \cos(ty/u_1))}{y} dy - \int_0^1 \frac{e^{\xi y/u_2} (1 - \cos(ty/u_2))}{y} dy \\
    &\geq \frac{1}{(\xi/u_1)^2 + \pi^2} + O(1).
\end{align*}
\] (3.19)

Indeed, we refer to the proof in [69] p. 374 that the first integral in Equation (3.19) is greater than or equal to \(u_1/((\xi/u_1)^2 + \pi^2)\). Next we give a lower bound for the second integral

\[
\int_0^1 \frac{e^{\xi y/u_2} (1 - \cos(ty/u_2))}{y} dy = \frac{t}{u_2} \int_0^1 \frac{1 - \cos(ty/u_2)}{ty/u_2} dy.
\] (3.20)

For the range \(u_1 < |t| < u_2\), and since \(0 < y < 1\), we have \(|ty/u_2| < 1\). For \(x < 1\), from the Taylor expansion of \(\cos(x)\) we have

\[
\cos(x) = 1 - O(x^2),
\]

which implies that

\[
\frac{1 - \cos(x)}{x} = O(x).
\]

Applying this to the integral in Equation (3.20), we obtain

\[
\int_0^1 \frac{e^{\xi y/u_2} (1 - \cos(ty/u_2))}{y} dy = O(1) \frac{t}{u_2} \int_0^1 \frac{ty}{u_2} e^{\xi y/u_2} dy
\]

\[
= O(1) \frac{t^2}{u_2^2} \int_0^1 y e^{\xi y/u_2} dy \leq O(1) \frac{t^2}{u_2^2} \int_0^1 e^{\xi y/u_2} dy
\]

\[
= O(1) \frac{t^2}{u_2^2} \frac{u_2}{\xi} (e^{\xi/u_2} - 1) = \frac{t^2}{u_2^2} \frac{u_2}{\xi} O \left( \frac{\xi}{u_2} \right) \leq O(1)
\]
where we used the assumption $u_1 \log u_1 = o(u_2)$, which implies that $\xi/u_2 \to 0$, and therefore $e^{\xi/u_2} - 1 = O(\xi/u_2)$. Also in this range $|t/u_2| < 1$ which proves Equation (3.19) and concludes the proof.

We break up the tails in three parts: $|t| \geq u_2$, $u_1 < |t| < u_2$ and $\delta < |t| \leq u_1$. For the first range, the contribution is negligible due to the above lemma, and the following easy facts:

$$u_1 = O\left( e^{\xi/u_2} - E\left(\frac{-\xi}{u_2}\right) - E\left(\frac{-\xi}{u_1}\right) \right),$$

and

$$f(-\xi) = -\xi + E\left(\frac{-\xi}{u_2}\right) - E\left(\frac{-\xi}{u_1}\right).$$

We concentrate now in the range $\delta < |t| \leq u_1$. By the above lemma, the tail in this range is upper bounded by

$$\int_{|t| < u_1} \exp\left(f(-\xi) - \frac{K}{u_1} t^2\right) \, dt \leq \exp(f(-\xi)) \int_{\delta}^{\infty} e^{-\frac{K}{u_1} t^2} \, dt$$

$$\leq \frac{\exp(f(-\xi)) \sqrt{u_1}}{2\sqrt{K}} \int_{\delta^2 K/u_1}^{\infty} e^{-y} \frac{dy}{\sqrt{y}}$$

$$\leq \frac{\exp(f(-\xi)) u_1}{2\delta K} e^{-K \delta^2 / u_1}$$

$$= \frac{\exp(f(-\xi)) \sqrt{u_1}}{2K \log(u_1)} e^{-K (\log u_1)^2}$$

which is clearly acceptable.

For the intermediate range $u_1 < |t| < u_2$, the tails are bounded by

$$\exp\left(e^{\xi/u_1} u_2 / n\right) \exp(f(-\xi)) e^{-\left(\log u_1\right)^2 / 2}(u_2 - u_1),$$

which again is acceptable provided that the above expression is $o(\sqrt{u_1})$. This holds by the hypothesis of Theorem 3.4.3 since $m_2 \gg \log^3 n$ and $m_2 \gg \sqrt{n m_1 (e^{-n/(m_1 \log n)}).}$ Putting all pieces together, we have proven that

$$N_q(n, m_1, m_2) = q^n e^{O(e^{\xi/u_1} (\xi + \delta)/n)} \int_{-\xi - i \infty}^{-\xi + i \infty} e^{E(h/u_2) - E(h/u_1)} e^h \, dh.$$

Under the condition $\sqrt{n \log n} \ll m_1$, we have

$$e^{O(e^{\xi/u_1} (\xi + \delta)/n)} = 1 + o(1).$$

We have proven the following theorem.
Theorem 3.4.3 The number \( N_q(n, m_1, m_2) \) of monic polynomials over \( \mathbb{F}_q \) with all irreducible factors with degree between \( m_1 \) and \( m_2 \), with \( m_1, m_2 \to \infty \) satisfying \( \sqrt{n} \log n \ll m_1 \ll n \) and \( \max\{\log^3 n, \sqrt{n m_1} e^{-n/(m_1 \log n)^2}\} \ll m_2 \ll m_1 / \log n \), is asymptotically

\[
N_q(n, m_1, m_2) \sim \frac{q^n}{2n\pi i} \int_{-\xi-i\infty}^{-\xi+i\infty} e^{E(h/u_2) - E(h/u_1)} e^h dh.
\]

where \( u_1 = n/m_1 \), \( u_2 = n/m_2 \).

The proof of Theorem 3.4.3 gives more than the integral form stated. The following corollary gives the asymptotic estimate obtained in the proof of Theorem 3.4.3. The estimate is stated in terms of the integral

\[
Ei(x) = \int_{-\infty}^{x} \frac{e^t}{t} dt \quad (x > 0).
\]

The reason for this is that \( Ei(x) \) is a real valued function, and the final result is more natural expressed in that way.

Corollary 3.4.4 The number \( N_q(n, m_1, m_2) \) of monic polynomials over \( \mathbb{F}_q \) with all irreducible factors with degree between \( m_1 \) and \( m_2 \), with \( m_1, m_2 \to \infty \), \( \sqrt{n} \log n \ll m_1 \ll n \) and \( \max\{\log^3 n, \sqrt{n m_1} e^{-n/(m_1 \log n)^2}\} \ll m_2 \ll m_1 / \log n \), is asymptotically

\[
N_q(n, m_1, m_2) \sim \frac{q^n \sqrt{m_1}}{2n\pi} \exp\left( Ei(\xi/u_1) - Ei(\xi/u_2) - \xi \right),
\]

where \( u_1 = n/m_1 \), \( u_2 = n/m_2 \).

Proof. From the proof of Theorem 3.4.3 we have

\[ N_q(n, m_1, m_2) \sim q^n J_3(n, m_1, m_2). \]

This combined with Equation (3.13) implies that

\[ N_q(n, m_1, m_2) \sim \frac{q^n \exp(f(-\xi))}{n\sqrt{2\pi}|f''(-\xi)|}. \]

As it was pointed out in Equation (3.12), \( f''(-\xi) \sim 1/u_1 \). Thus, it only remains to estimate \( f(-\xi) \)

\[
f(-\xi) = E(-\xi/u_2) - E(-\xi/u_1) - \xi
= -Ei(\xi/u_2) + Ei(\xi/u_1) - \xi,
\]

where the second equality holds since \( E(x + i0) = -Ei(-x) - i\pi \) (see [1], 5.1.7).

Again, as for the generalized smooth polynomials, one can extend the range of the estimate considerably, by weakening the result.
Theorem 3.4.5 The number \( N_q(n, m_1, m_2) \) of monic polynomials over \( \mathbb{F}_q \) with all irreducible factors between \( m_1 \) and \( m_2 \), with \( m_1, m_2 \to \infty \), \( m_1 e^{-n/m_1} \ll m_2 \leq cm_1 \) for any constant \( c < 1 \), and \( 2(\log n)^2 \leq m_1 \ll n \) satisfies

\[
N_q(n, m_1, m_2) = q^n e^{-(1+o(1))u_1 \log u_1},
\]

where \( u_1 = n/m_1 \).

Proof. The number \( N_q(n, m_1, m_2) \) was estimated in terms of four integrals that correspond to the ranges \( |t| \leq \delta, \delta < |t| \leq u_1, u_1 < |t| < u_2 \), and \( |t| \geq u_2 \). From the proof of Theorem 3.4.3, it is clear that the main integral corresponding to the range \( |t| \leq \delta \) is \( e^{-(1+o(1))u_1 \log u_1} \) with the only assumption that \( \log n \ll m_1 \ll n \). Moreover, the tail integrals that correspond to the ranges \( |t| \geq u_2 \) and \( \delta < |t| \leq u_1 \) are \( e^{-(1+o(1))u_1 \log u_1} \) under no further assumption. The rest of the conditions come from the range \( u_1 < |t| < u_2 \). In that range the tail is

\[
\exp\left(\frac{e^{\xi u_1} u_2}{n}\right) \exp(f(-\xi)) e^{\frac{u_1}{n} u_2} = \exp(f(-\xi)) e^{\frac{u_1}{n} u_2 + \frac{u_1 u_2 \log u_1 + \log u_2}{n}}.
\]

We know that \( f(-\xi) = -(1 + o(1))u_1 \log u_1 \), and all the other terms in the exponent are \( o(1)u_1 \log u_1 \), except maybe \( \log u_2 \). In order to ensure this, we need to impose the condition \( m_2 \gg m_1 e^{-n/m_1} \). The theorem now follows.

In the next chapter, we use the results proved here to analyze a generalized version of the index calculus method for the discrete logarithm problem in finite fields.
Chapter 4

The discrete logarithm problem in finite fields

4.1 Introduction

In this chapter, we study the problem of computing discrete logarithms in the multiplicative group of certain finite fields. In particular, we consider finite fields $\mathbb{F}_p^n$, where the characteristic $p$ is small (a fixed constant), and the extension degree $n$ is large (formally it tends to infinity). We represent the elements of $\mathbb{F}_p^n$ as polynomials over the base field $\mathbb{F}_p$ modulo a monic irreducible of degree $n$. This representation is possible because of the isomorphism

$$\mathbb{F}_p^n \cong \frac{\mathbb{F}_p[x]}{(f)},$$

where $f$ is a monic irreducible polynomial over $\mathbb{F}_p[x]$ of degree $n$.

For the DLP in finite fields of small characteristic the index calculus method has been provably successful: its running time has been proven to be subexponential. In that sense the algorithm can be considered well-understood. However, there is a very important parameter, namely the form of the “factor base”, that was never questioned. The goal of this chapter is to study the behavior of the index calculus method under different factor bases. To achieve that, in Section 4.2, we give a description of the algorithm with the factor base being any set $S$ of monic irreducible polynomials in $\mathbb{F}_p[x]$. Then, in Section 4.3.1, we compute the size of $S$. The index calculus method is a randomized algorithm, so its running time depends on the probability that certain steps are successful. In Section 4.3.2 we determine the probability of success. It is this part of the analysis that uses the density results of Chapter 3. The estimation of the time complexity of the algorithm then follows. The conclusion of the theoretical analysis is that the running
time of the algorithm remains unaffected under a wide range of different factor bases. An interesting (and important) property of the change we introduce is that it does not "interfere" with the algebraic manipulations of the method. Several variants of the basic method have been introduced, all of them using the standard factor base, which aim to speed up the algorithm by changing the algebraic manipulations. Two of the earliest variants are the Waterloo method, found by Blake, Fuji-Hara, Mullin, and Vanstone [7], and the variant of Coppersmith [13]. In Section 4.4, we consider those variants under a general factor base and show that their running time remains unaffected as well. Finally, in Section 4.5 we report on some experimental results. The results of this chapter can be found in [25].

4.2 The algorithm

Let \( q = p^n \), for a prime \( p \) and an integer \( n > 1 \). We start by describing the index calculus method for computing discrete logarithms in \( \mathbb{F}_q \). The elements in \( \mathbb{F}_q \) are represented as polynomials over \( \mathbb{F}_p \) of degree smaller than \( n \). The arithmetic is modulo a monic irreducible polynomial \( f \in \mathbb{F}_p[x] \) of degree \( n \). We are given \( p, n \) that determine the finite field, \( f \) that determines the arithmetic, a generator \( g \) of the multiplicative group of \( \mathbb{F}_p[x]/(f) \), and the polynomial \( h^* \in \mathbb{F}_p[x] \) whose discrete logarithm we want to compute. Then we choose a set \( S \) of monic irreducible polynomials over \( \mathbb{F}_p \). We call \( S \) the factor base. The method consists of two stages.

Algorithm: Index Calculus

Input: A prime \( p \), an integer \( n \), a monic irreducible polynomial \( f \) of degree \( n \), a polynomial \( h^* \), and the generator \( g \).

Output: An integer \( \ell \) such that \( h^* = g^\ell \).

Stage 1

1. Choose an integer \( s \) in \([1, q - 1]\) uniformly at random, and form the polynomial 
   \( h \equiv g^s (\text{mod } f) \), \( \deg h < n \).

2. Check if \( h \) factors completely into irreducibles over \( S \). If not, discard it. If it does, say

   \[ h = \prod_{v \in S} v^{v(h)}, \]
record the congruence
\[ s \equiv \sum_{v \in S} e_v(h) \log_g v \pmod{q - 1}. \]

Repeat steps 1 and 2 until "slightly more" than \#S congruences are obtained.

3. Solve the system to determine \( \log_g v \) for all \( v \in S \).

**Stage 2**

1. Choose an integer \( s \) in \([1, q - 1]\) uniformly at random; form the polynomial \( h \equiv h \cdot g^s \pmod{f}, \deg h < n \).

2. Check if \( h \) factors completely into irreducibles over \( S \). If not, discard it. If it does, say
\[
    h = \prod_{v \in S} v^{e_v(h)},
\]
compute the required discrete logarithm as
\[
    \ell = \log_g h^* \equiv -s + \sum_{v \in S} e_v(h) \log_g v \pmod{q - 1}.
\]

The method works for any factor base \( S \). The choice of the factor base however, clearly affects the time complexity of the algorithm. For example, a "very large" factor base, that would speed up the second stage, would make the first stage (and thus the whole method) totally inefficient. Until now, it was generally assumed that the optimal choice of a factor base was the set of all monic irreducible polynomials of degrees smaller than or equal to \( m \) for a certain parameter \( m \), the smooth factor base, although this assumption has not been established as a theorem.

We consider a more general version, when the factor base consists of all monic irreducibles of degree between \( m_2 \) and \( m_1 \). Our analysis shows that the important parameter here is \( m_1 \). In the next section, we show that the asymptotic running time of the algorithm remains the same as that of the basic version even if \( m_2 \) is of the same order as \( m_1 \). This means that the size of the factor base could be smaller, an interesting fact for practical purposes when space is a constraint.

### 4.3 Analysis

The analysis of the algorithm in Section 4.2 requires the study of the size of the factor base and the probability of successfully factoring polynomials into irreducibles over \( S \).
4.3.1 Size of the factor base

We start by estimating the size \( t \) of the factor base.

**Proposition 4.3.1** Let \( S \) be a factor base formed by irreducible polynomials over \( \mathbb{F}_p \) with degree between \( m_2 \) and \( m_1, m_2 < m_1 \). Then, as \( m_1 \to \infty \), the size \( t \) of the factor base \( S \) is upper bounded by

\[
e^{(1+o(1))m_1 \log p}.
\]

**Proof.** It is well-known (see for instance [36], Theorem 3.25) that the number \( I_k \) of monic irreducible polynomials of degree \( k \) over \( \mathbb{F}_p \) is given by

\[
I_k = \frac{1}{k} \sum_{d \mid k} \mu(d) p^{k/d} = \frac{p^k}{k} + \sum_{d \mid k, \ d > 1} \mu(d) p^{k/d}.
\]

It is easy to derive the following upper bound

\[
\left| I_k - \frac{p^k}{k} \right| = \left| \sum_{d \mid k, \ d > 1} \mu(d) p^{k/d} \right| \leq \sum_{j=1}^{\lfloor k/2 \rfloor} p^j = \frac{p^{\lfloor k/2 \rfloor + 1} - 1}{p - 1} < 2p^{k/2}.
\]

An upper bound on the size of the factor base can be computed as \( m_1 \to \infty \)

\[
t = \sum_{k=m_2+1}^{m_1} I_k < \sum_{k=m_2+1}^{m_1} \left( \frac{p^k}{k} + 2p^{k/2} \right)
\]

\[
\leq \frac{1}{m_2} \sum_{k=m_2+1}^{m_1} p^k + 2 \sum_{k=m_2+1}^{m_1} p^{k/2} = \frac{1}{m_2} \left( \frac{p^{m_2+1} - p^{m_2+1}}{p - 1} \right) + O \left( p^{m_1/2} \right)
\]

\[
\leq \frac{p^{m_1+1}}{m_2} + O \left( p^{m_1/2} \right) = e^{(1+o(1))m_1 \log p}.
\]

\( \square \)

The quantity needed for the analysis of the algorithm is the size of the system of congruences created in the first stage. It is shown in [37] that if \( 4t \log p^n \) linear congruences are computed, the probability that the system has full rank is at least \( 1/2 \).

4.3.2 Probability of success

We now give an estimate on the number of repetitions needed in both stages until the polynomial \( h \) completely factors over the factor base \( S \).

As it will be clear below, we need an estimate for the probability that a random monic polynomial of degree at most \( n-1 \) factors over \( S \), i.e., it factors into irreducibles
of degree greater than \( m_2 \), and less than or equal to \( m_1 \). This probability is determined using Theorems 3.3.2 and 3.4.5. That is the reason we need to assume that \( m_1 \) and \( m_2 \) satisfy the assumption of those theorems. Namely, we require that \( \log n \ll m_1 \ll n \) and either \( m_2 \) is constant, or \( m_1 e^{-n/m_1} \ll m_2 \leq cm_1 \) for any constant \( c < 1 \). The condition on \( m_1 \) is not restrictive at all, since the optimal choice will turn out to be \( c' \sqrt{n \log n} \), i.e., within the range. For this choice of \( m_1 \), \( m_2 \) is allowed to be either a constant or any function of \( n \), provided of course that \( m_2 = O(m_1) \), since \( m_1 e^{-n/m_1} \rightarrow 0 \).

The next proposition gives an upper bound on the number of repetitions needed in both stages until the polynomial \( h \) completely factors over the factor base \( S \).

**Proposition 4.3.2** Let \( S \) be a factor base formed by irreducible polynomials over \( \mathbb{F}_p \) with degree between \( m_2 \) and \( m_1 \), with \( m_2 \) and \( m_1 \) as in the hypothesis of either Theorem 3.3.2 or Theorem 3.4.5. Then, the expected number of repetitions of Steps (1) and (2) in both Stage 1 and Stage 2 of the algorithm in Section 4.2 is upper bounded by

\[
e^{(1+o(1)) \frac{n-1}{m_1} \log \frac{n}{m_1}}. \tag{4.2}
\]

**Proof.** We recall that the polynomial \( h \) in the algorithm of Section 4.2 is a random monic polynomial in \( \mathbb{F}_p[x] \) of degree at most \( n - 1 \), and the experiment is a sequence of Bernoulli trials. The expected number of repetitions until we have the first success is one over the probability of success. Thus, a lower bound on the probability of success would provide an upper bound on the expected number of repetitions. Then, we have

\[
Pr(h \text{ factors over } S) = \sum_{k=1}^{n-1} Pr(h \text{ factors over } S| \deg h = k) Pr(\deg h = k)
\]

\[
\geq Pr(h \text{ factors over } S| \deg h = n - 1) Pr(\deg h = n - 1).
\]

It is easy to see that the probability of a random monic polynomial having degree \( n - 1 \) is very close to 1. More precisely, we have

\[
Pr(\deg h = n - 1) = \frac{\# \text{ polynomials of degree } n - 1}{\# \text{ polynomials of degree } \leq n - 1} = \frac{p^{n-1}}{\sum_{i=1}^{n-1} p^i}
\]

\[
= \frac{p^{n-1}(p-1)}{p^n - p} = 1 - \frac{p^{n-1} - p}{p^n - p} > 1 - \frac{1}{p}.
\]

It remains to combine this with the probability that a random monic polynomial of degree \( n - 1 \) factors over \( S \) given in Theorems 3.3.2 and 3.4.5. Choosing the parameters \( m_1 \) and \( m_2 \) so that they satisfy the conditions of these theorems, we can estimate the probability of interest as

\[
Pr(h \text{ factors over } S| \deg h = n - 1) = e^{-(1+o(1))u_1 \log u_1},
\]
where \( u_1 = n/m_1 \). In fact, one should let \( u_1 = (n - 1)/m_1 \), but the \( o(1) \) in the exponent "neutralizes" such small changes. Therefore, the expected number of repetitions is upper bounded by

\[
e^{(1+o(1)) \frac{n}{m_1} \log \frac{n}{m_1}},
\]

where the multiplicative term \( (1 - 1/p)^{-1} \) due to the probability that \( h \) is of degree \( n - 1 \) is absorbed by the \( o(1) \) in the exponent.

4.3.3 Time complexity

We have all the quantities needed to analyze the algorithm. In this section we will compute a precise value for \( m_1 \), in terms of \( n \), and give an upper bound for the running time of the index calculus method in terms of the function

\[
L(n) = e^{(1+o(1)) \sqrt{n \log n}}.
\]

The main result (Theorem 4.3.3) shows that asymptotically our factor base is as good as the standard base used in the basic index calculus algorithm.

As it is the case when the standard factor base is used, the running time of the algorithm is dominated by the first stage. This is when a tradeoff takes place regarding the size of the factor base: large \#S means small number of repetitions (until a useful congruence is found), but many such congruences are needed for the system to be solvable. The complexity of the algorithm is proven in the following theorem.

Theorem 4.3.3 The running time of the algorithm in Section 4.2 is, as \( n \to \infty \),

\[
L(n)^{\sqrt{2 \log p}}.
\]

Proof. From Equations (4.1) and (4.2) and the fact that \( 4tn \log p \) congruences have to be generated, where \( t \) is the size of the factor base, we conclude that the time to create the system is

\[
4n \log p e^{(1+o(1)) \left( m_1 \log p + \frac{n}{m_1} \log \frac{n}{m_1} \right)} \sim e^{(1+o(1)) \left( m_1 \log p + \frac{n}{m_1} \log \frac{n}{m_1} \right)}.
\]

Furthermore, if a method for sparse linear systems is used like the method proposed by Wiedemann [70] (see also [48]), then the time for solving the system is

\[
(4tn \log p)^2 \sim e^{(2+o(1))(m_1 \log p)}.
\]
Thus, the asymptotic running time of the first stage is given by
\[
e^{(1+o(1))(m_1 \log p + \frac{n}{m_1} \log \frac{n}{m_1})} + e^{(2+o(1))(m_1 \log p)}.
\] (4.3)

We note here that the computation of \( g^x \mod f \) is done by repeated squaring, and takes time polynomial in \( n \) and \( \log p \). Moreover, the factorization of the polynomials can be done in probabilistic polynomial time (see [26] for a recent survey on the topic). Those computations introduce a multiplicative polynomial factor in the above estimate, which is absorbed in the \( o(1) \) of the exponent.

Let us consider now \( m_1 = cn^\alpha (\log n)^\beta \), for some positive constants \( \alpha, \beta, c \) to be determined later. The first exponent in the above expression becomes
\[
(1 + o(1)) \left( cn^\alpha (\log n)^\beta \log p + \frac{1}{c} n^{1-\alpha} (\log n)^{2-\beta} \log \left( \frac{1}{c} n^{1-\alpha} (\log n)^{2-\beta} \right) \right),
\]
while the second exponent is
\[
(2 + o(1)) cn^\alpha (\log n)^\beta \log p.
\]

The expressions are minimized for \( \alpha = \beta = \frac{1}{2} \). After the substitutions, the expressions become
\[
(1 + o(1)) \left( c \log p + \frac{1}{2c} \right) \sqrt{n \log n},
\]
and
\[
(1 + o(1)) \left( 2c \log p \sqrt{n \log n} \right).
\]

We view the multiplicative constant \( c \log p + \frac{1}{2c} \) as a function of \( c \), and observe that it achieves a minimum at \( c = (2 \log p)^{-1/2} \). This minimum is \( \sqrt{2 \log p} \). Substituting everything in Equation (4.3), we obtain an upper bound for the running time of the first stage
\[
L(n)^{2 \log p} = e^{(\sqrt{2 \log p} + o(1)) \sqrt{n \log n}}.
\]

Now that the parameters have been fixed, we can easily compute an upper bound for the second stage. This is the expected number of repetitions until the first success. Therefore, an upper bound is
\[
e^{(1+o(1))m_1 \log p} = e^{(1+o(1)) \log p(2 \log p)^{-1/2} \sqrt{n \log n}} = e^{\left( \sqrt{\log p/2} + o(1) \right) \sqrt{n \log n}} = L(n)^{\sqrt{\log p/2}}.
\]

Clearly, the running time of the algorithm is dominated by the first stage, and an upper bound for it is, as \( n \to \infty \),
\[
L(n)^{\sqrt{2 \log p}}.
\]
**Remark.** The algorithm as described in Section 4.2 works with any factor base $S$. The subsequent analysis shows that if the factor base $S$ consists of all monic irreducible polynomials over $\mathbb{F}_p$ with degree greater than $m_2$ and less than or equal to $\sqrt{n \log n/(2 \log p)}$, for any $0 \leq m_2 \leq c \sqrt{n \log n/(2 \log p)}$ then the algorithm has running time $L(n)^{\sqrt{2 \log p}}$. The reason is that the size of the factor base and the probability of success depend only on $m_1$ — in fact they also depend on $m_2$, but this dependence is very weak, and is hidden in the $o(1)$ in the exponent. It is clear now, that if we choose the factor base to consist of the set of all monic irreducibles having degree in a set $T$, then the running time will remain the same provided that $T$ contains all integers from $c \sqrt{n \log n/(2 \log p)}$ to $\sqrt{n \log n/(2 \log p)}$, for any constant $c < 1$. In other words, $T$ does not necessarily have to contain consecutive integers, provided that the above condition is met. The situation is shown in Figure 4.1.
4.4 Some remarks

4.4.1 The Waterloo variant

We now turn to a variant of the basic method in Section 4.2, that was proposed by Blake, Fuji-Hara, Mullin, and Vanstone [7], known as the Waterloo variant. It is one of the first variants that appeared in the literature, and attempts to improve the running time of the method at the cost of not being able to analyze the variant rigorously. In this section, we show that the same heuristic arguments go through for the generalized factor base.

In the Waterloo variant, the following change is introduced: in step 2 of both stages, instead of trying to factor the polynomial \( h \) one tries to find two polynomials \( w_1 \) and \( w_2 \) of degree at most \( n/2 \) each, such that

\[
w_1 h \equiv w_2 (\text{mod } f),
\]

and then tries to factor them over the factor base \( S \). Such polynomials \( w_1 \) and \( w_2 \) can be computed easily using the extended Euclidean algorithm on input \( (h, f) \). If one succeeds and say

\[
w_i = \prod_{v \in S} v^{e_v(w_i)}, \quad i = 1, 2
\]

then the congruence

\[
s \equiv \sum_{v \in S} (e_v(w_2) - e_v(w_1)) \log v (\text{mod } q - 1)
\]

is recorded. The algorithm otherwise remains the same. More details about the method are given in the original paper [7], and in the survey article by Odlyzko [48].

The problem for a rigorous analysis is that not much is known about the joint distribution of \( (w_1, w_2) \). The heuristic assumption is that \( w_1 \) and \( w_2 \) behave as random independent polynomials of degree at most \( n/2 \). The probability that two random monic independent polynomials of degree at most \( n/2 \) factor into irreducibles of degree greater than \( m_2 \) and at most \( m_1 \) is

\[
e^{-2(1+o(1)) \frac{n}{2m_2} \log \frac{n}{2m_2}} = e^{-(1+o(1)) \frac{n}{m_1} \log \frac{n}{2m_1}}. \tag{4.4}
\]

The size of the factor base remains the same as in the basic version, therefore an upper bound on the running time for the first stage is

\[
e^{(1+o(1)) (m_1 \log p + \frac{n}{m_1} \log \frac{n}{2m_1})} + e^{(2+o(1)) m_1 \log p}, \tag{4.5}
\]
where we consider the same algorithms and costs as in Theorem 4.3.3. Again we let 
\( m_1 = c \sqrt{n \log n} \), and Equation (4.5) becomes
\[
\sum_{i=1}^{d} \left( 1 + o(1) \right) \left( c \log p + \frac{1}{2c} \right) \sqrt{n \log n} + \sum_{i=1}^{d} \left( 1 + o(1) \right) \left( 2c \log p \right) \sqrt{n \log n}.
\]
The above expression is minimized for \( c = (2 \log p)^{-1/2} \), and the running time of the first stage is bounded by
\[
L(n)^{\frac{1}{2} \log p}.
\]
For this choice of parameters, it is easy to see that the second stage has running time bounded by
\[
L(n)^{\frac{1}{2} \log p / 2}.
\]
Therefore, the running time of the Waterloo variant is bounded by
\[
L(n)^{\frac{1}{2} \log p}.
\]

**Remark.** The practical improvement that this variant provides is hidden in the above analysis in the \( o(1) \) term of the exponent. To see the actual improvement one should compare Equation (4.3) with Equation (4.5), and in particular the time to create the system of congruences (the two systems are expected to have the same size, and therefore the time to solve it is expected to be of the same order). The Waterloo variant is faster by a factor of \( 2^{n/m_1} \), as was the case when the standard base was used (see [48], pp. 238-243).

### 4.4.2 The Coppersmith variant

In this section, we briefly consider the variant proposed by Coppersmith in [13]. It is designed to work in finite fields of characteristic two, and it relies on unproven assumptions. which however, seem to be reasonable (at least in practice). We note that this was the first method to be conjectured to have running time of the form
\[
M(n) = e^{\left( c + o(1) \right) n^{1/3} (\log n)^{2/3}},
\]
where \( c \) is an effectively computable constant.

The method assumes that the monic irreducible polynomial \( f \) of degree \( n \) used to define the field \( \mathbb{F}_{2^n} \) is of the form \( f(x) = x^n + f_1(x) \), where \( \deg f_1 \leq \log_2 n \). Little is known about the irreducibility of this type of polynomials; see [23] for general computational experiments with sparse polynomials over \( \mathbb{F}_2 \) and for applications of these polynomials. There exist some values of \( n \) with no irreducible polynomial of this form. On the other hand, if we let \( \deg f_1 \leq \log_2 n + c_1 \), for \( c_1 \) a small positive integer, there are irreducible polynomials of this form of degree \( n \) for all practical values of \( n \) (see [23]).
In the first stage of the algorithm, two polynomials $w_1$ and $w_2$ are constructed such that
\[ w_2 \equiv w_1^{2^k} \pmod{f}, \tag{4.6} \]
where $k$ is a parameter chosen so that $2^k$ is on the order of $n^{1/3}(\log n)^{-1/3}$. The construction of $w_1$ and $w_2$ is based on algebraic manipulations that are independent of the factor base. Next, the heuristic assumption is made that $w_1$ and $w_2$ behave like random independent polynomials of degree at most on the order of $n^{2/3}(\log n)^{1/3}$. The analysis follows by essentially the same arguments given in the previous sections. The analysis of the second stage is carried out in a similar manner. For a detailed description of the algorithm, the reader is referred to the original paper by Coppersmith [13], and to the excellent survey by Odlyzko [48].

### 4.5 Experimental results

In this section we present some experimental results for the quantities of interest for the analysis of the basic index calculus method. We fix the degree of the field extension to $n = 500$, and consider three different values for $p$, namely 2, 3, and 5. Having fixed the values for $n$ and $p$, we compute the corresponding value for $m_1$ as determined in Section 4.3.3. that is
\[ m_1 = \sqrt[2]{\frac{n \log n}{2 \log p}}. \]

We consider several possible values for $m_2$, since this is a "free parameter". The value $m_2 = 0$ corresponds to the standard method with the smooth factor base. For each value of $m_2$ we compute the number of polynomials of interest $N_p(n, m_1, m_2)$, the probability of success computed as $p^{-n}N_p(n, m_1, m_2)$, and the size $\#S$ of the factor base.

About the finite fields of the experiments, we comment that in Table 1 the size of the field is $2^{500} \approx 10^{150}$, which is on the borderline of what is currently computable. The finite fields in the Tables 2 and 3 are of size $3^{500} \approx 10^{239}$ and $5^{500} \approx 10^{349}$ respectively, which is well beyond the size of the fields for which discrete logarithms are currently computable.

The experiments indicate that for moderately large (but quite reasonable for practical purposes) values for $n$ and $p$, the probability of success drops faster than the size of the factor base. This suggests that the common belief that the smooth factor base is "optimal" is, practically speaking, justified.
Table 4.1: $n = 500$, $p = 2$, $m_1 = 47$.

<table>
<thead>
<tr>
<th>$m_2$</th>
<th>$N(n, m_1, m_2)$</th>
<th>Prob</th>
<th>#S</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>.1240e140</td>
<td>.1240e-10</td>
<td>.6125243528e13</td>
</tr>
<tr>
<td>1</td>
<td>.2613e139</td>
<td>.2613e-11</td>
<td>.6125243528e13</td>
</tr>
<tr>
<td>4</td>
<td>.1001e139</td>
<td>.1001e-11</td>
<td>.6125243528e13</td>
</tr>
<tr>
<td>8</td>
<td>.3755e138</td>
<td>.3755e-12</td>
<td>.6125243528e13</td>
</tr>
<tr>
<td>12</td>
<td>.1579e138</td>
<td>.1579e-12</td>
<td>.6125243527e13</td>
</tr>
<tr>
<td>15</td>
<td>.8222e137</td>
<td>.8222e-13</td>
<td>.6125243523e13</td>
</tr>
<tr>
<td>23</td>
<td>.1116e137</td>
<td>.1116e-13</td>
<td>.6125242762e13</td>
</tr>
</tbody>
</table>

Table 4.2: $n = 500$, $p = 3$, $m_1 = 38$.

<table>
<thead>
<tr>
<th>$m_2$</th>
<th>$N(n, m_1, m_2)$</th>
<th>Prob</th>
<th>#S</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>.1975e224</td>
<td>.1975e-15</td>
<td>.5406460513e17</td>
</tr>
<tr>
<td>1</td>
<td>.2471e223</td>
<td>.2471e-16</td>
<td>.5406460513e17</td>
</tr>
<tr>
<td>3</td>
<td>.1049e223</td>
<td>.1049e-16</td>
<td>.5406460513e17</td>
</tr>
<tr>
<td>5</td>
<td>.5347e222</td>
<td>.5347e-17</td>
<td>.5406460513e17</td>
</tr>
<tr>
<td>8</td>
<td>.2220e222</td>
<td>.2220e-17</td>
<td>.5406460513e17</td>
</tr>
<tr>
<td>15</td>
<td>.2742e221</td>
<td>.2742e-18</td>
<td>.5406460512e17</td>
</tr>
<tr>
<td>18</td>
<td>.9309e220</td>
<td>.9309e-19</td>
<td>.5406460509e17</td>
</tr>
</tbody>
</table>

Table 4.3: $n = 500$, $p = 5$, $m_1 = 31$.

<table>
<thead>
<tr>
<th>$m_2$</th>
<th>$N(n, m_1, m_2)$</th>
<th>Prob</th>
<th>#S</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>.1297e331</td>
<td>.1297e-18</td>
<td>.1893590989e21</td>
</tr>
<tr>
<td>1</td>
<td>.1473e329</td>
<td>.1473e-20</td>
<td>.1893590989e21</td>
</tr>
<tr>
<td>3</td>
<td>.5304e328</td>
<td>.5304e-21</td>
<td>.1893590989e21</td>
</tr>
<tr>
<td>5</td>
<td>.2369e328</td>
<td>.2369e-21</td>
<td>.1893590989e21</td>
</tr>
<tr>
<td>8</td>
<td>.7794e327</td>
<td>.7794e-22</td>
<td>.1893590989e21</td>
</tr>
<tr>
<td>12</td>
<td>.1561e327</td>
<td>.1561e-22</td>
<td>.1893590989e21</td>
</tr>
<tr>
<td>15</td>
<td>.3549e326</td>
<td>.3549e-23</td>
<td>.1893590989e21</td>
</tr>
</tbody>
</table>
Chapter 5

Basic notions on elliptic curves

In this chapter, we give a brief introduction to the theory of elliptic curves over finite fields. As the theory of elliptic curves is both extensive and deep, a self-contained and complete introduction to the topic is virtually impossible. Our intention is to introduce the basic concepts in a way that requires no more than elementary algebra. Most of the theorems and propositions will be given without proofs. The reader is referred to the book by Silverman [64] for an extensive treatise on the subject, and the technical report by Charlap and Robbins [12] for an elementary presentation.

5.1 Elliptic curves

For the rest of the chapter, $K$ will denote a field. The characteristic of a field $K$ is the smallest positive integer $p$, such that $p \cdot 1 = 0$, where by $p \cdot 1$ we mean $1 + 1 + \cdots + 1$, $p$ times. It can be shown that if such a $p$ exists, then it is always a prime. If no such $p$ exists, then we say that the characteristic is zero. A field is said to be algebraically closed if every polynomial of degree $n$ with coefficients in the field has $n$ roots in the field; that is, it splits into linear factors with coefficient in the field. The standard example of an algebraically closed field is $\mathbb{C}$, the field of complex numbers. It is well known, that every field has a minimal algebraically closed superfield, called its algebraic closure. We denote the algebraic closure of $K$ by $\overline{K}$. We note, that the algebraic closure of a field is an infinite set, even if the base field is a finite field. For instance, if $K = \mathbb{F}_q$, the finite field with $q$ elements, then its algebraic closure $\overline{\mathbb{F}_q}$, can be thought as the union of $\mathbb{F}_{q^k}$ for $k = 1, 2, \ldots$. 

49
We consider polynomials in $K[X, Y]$, i.e., expressions of the form
\[ C(X, Y) = \sum_{i,j} a_{ij} X^i Y^j, \quad a_{ij} \in K. \]

The degree of a term $a_{ij} X^i Y^j$ is defined to be $i + j$. Then the degree of $C$ is the maximum of the degrees of its terms. A root or zero of $C$ is a point $P = (a, b) \in \overline{K} \times \overline{K}$ such that $C(P) = 0$.

**Definition 5.1.1** A root $P \in \overline{K} \times \overline{K}$ of a polynomial $C \in K[X, Y]$ is said to be a singular point if it satisfies
\[ \frac{\partial C}{\partial X}(P) = \frac{\partial C}{\partial Y}(P) = 0. \]

It can be proven that every elliptic curve is isomorphic to some elliptic curve defined by a Weierstrass equation (with no singular points). Thus every elliptic curve has a model given by a Weierstrass equation.

**Definition 5.1.2** An elliptic curve $E$ over $K$, defined by a Weierstrass equation
\[ E : Y^2 + a_1 XY + a_3 Y = X^3 + a_2 X^2 + a_4 X + a_6, \quad (5.1) \]
with $a_1, ..., a_6 \in K$, with no singular points, is the set of all points in $\overline{K} \times \overline{K}$ that satisfy $E$, together with a special point $O$, called the point at infinity.

We write $E/K$ whenever an elliptic curve $E$ is defined by a Weierstrass equation with coefficients in $K$. We stress, that the point at infinity $O$, is not an element of $\overline{K} \times \overline{K}$. For our purposes, it can be thought simply as a special point, whose properties will be discussed later in this chapter. We note that depending on the characteristic of the field $K$, the defining equation may be further simplified. It should also be mentioned here that the degree that we used to define elliptic curves is not to be confused with the degree of a polynomial on $E$, which is defined in Section 5.2.

So far, we considered all the points with coordinates in $\overline{K}$. It makes perfect sense to consider the set of points in $K \times K$ that satisfy $C(X, Y) = 0$. This set, together with the point at infinity $O$ is called the set of points on $E$ defined over $K$, and is denoted by $E(K)$.

### 5.2 Polynomial and rational functions

The notion of polynomial and rational functions on curves is central in the theory of algebraic curves. Let $E$ be an elliptic curve defined by a polynomial equation $C(X, Y) =$
0. We would like to think of the elements in $\overline{K}[X, Y]$ as defining polynomial functions on $E$. Two elements in $\overline{K}[X, Y]$ should be considered the same, if they differ by a multiple of $C(X, Y)$. We call the set of polynomial functions on $E$, the coordinate ring of $E$.

**Definition 5.2.1** Let $E/K$ be an elliptic curve. The coordinate ring of $E/K$ is defined by

$$K[E] = \frac{K[X, Y]}{(C(X, Y))}.$$ 

**Remark.** Even in the case that the curve is defined over $K$, it makes sense to define the ring

$$\overline{K}[E] = \frac{\overline{K}[X, Y]}{(C(X, Y))}.$$ 

The symbols "$x$" and "$y$" are reserved for the coordinate functions on $E$ defined by $x(a, b) = a$ and $y(a, b) = b$. An alternative definition of the coordinate ring of $E$ is simply as $K[E] = K[x, y]$, since the functions $x$ and $y$ clearly satisfy $C$, and therefore $C(x, y) = 0$.

**Definition 5.2.2** A polynomial on $E$ is an element of $\overline{K}[x, y]$.

A consequence of this definition is that every polynomial on $E$ can be written as $f(x, y) = v(x) + yw(x)$, for two polynomials $v, w$ in one variable. We call this the canonical form of $f$. Then $f$ can be viewed as a function on $E$, since $x$ and $y$ are. We note that this representation of $f$ is unique.

The field of rational functions on $E$ is defined as the field of quotients of the coordinate ring.

**Definition 5.2.3** A rational function on $E$ is an equivalence class of formal quotients of polynomials $f/g$ ($g$ not identically zero), where we identify $f/g$ with $h/s$ if $f s = gh$ as polynomials on $E$. The set of rational functions on $E$ is a field denoted by $\overline{K}(E)$.

The field $\overline{K}(E)$ is defined similarly. We are interested in the value of rational functions on $E$.

**Definition 5.2.4** If $r$ is a rational function on $E$ and $P$ is a finite point in $E$, we say that $r$ is finite at $P$ if there exists a representation $r = f/g$, where $f$ and $g$ are polynomials on $E$ and $g(P) \neq 0$. If $r$ is finite at $P$, we define $r(P) = f(P)/g(P)$.

We note that if a rational function is finite at a finite point $P$ then the value $r(P)$ is well defined, i.e., it does not depend on the choice of $f$ and $g$.

The value of a rational function at the point at infinity is more complicated to define. First we define the degree of a polynomial on $E$. 

Definition 5.2.5 Let \( f(x, y) = v(x) + yw(x) \) be a non-zero polynomial on \( E \). We define the degree of \( f \) by

\[
\deg(f) = \max\{2 \cdot \deg_x(v), 3 + 2 \cdot \deg_x(w)\},
\]

where \( \deg_x \) is the usual degree of a univariate polynomial.

One way to justify this definition is to assign degree 2 to \( x \) and degree 3 to \( y \). Then

\[
2 \cdot \deg_x(v) \text{ is the degree of } v \quad \text{and} \quad 3 + 2 \cdot \deg_x(w) \text{ is the degree of } w.
\]

The maximum of the two is defined to be the degree of \( f \). We note that the degree map \( \deg : K[E] \to \mathbb{Z} \) defined here has the following property that one would expect:

\[
\text{for all non-zero } f, g \in K[E], \quad \deg(f \cdot g) = \deg(f) + \deg(g).
\]

We define now the value of a rational function at the point of infinity.

Definition 5.2.6 Let \( r = f/g \) be a rational function on \( E \). If \( \deg(f) < \deg(g) \), we set \( r(O) = 0 \). If \( \deg(f) > \deg(g) \), we say that \( r \) is not finite at \( O \). If \( \deg(f) = \deg(g) \), then writing \( f \) and \( g \) in canonical form, the leading terms (the terms of highest degree according to Definition 5.2.5) will be either \( ax^d \) and \( bx^d \) respectively (if \( \deg(f) \) is even) or \( ay^d \) and \( bx^d \) respectively (if \( \deg(f) \) is odd), for some \( a, b \in \overline{K} \) and some integer \( d \).

We define \( r(O) = a/b \).

If a rational function \( r \) is not finite at some point \( P \) (finite or \( O \)) we write \( r(P) = \infty \).

5.3 Zeros and poles

The definition of zeros and poles of rational functions is straightforward.

Definition 5.3.1 Let \( r \) be a rational function on \( E \). We say that \( r \) has a zero at \( P \in E \) if \( r(P) = 0 \) and that \( r \) has a pole at \( P \in E \) if \( r(P) = \infty \).

Next we want to define the multiplicity of a zero or pole. This is more complicated, and we need the following theorem.

Theorem 5.3.2 For each point \( P \in E \) there is a rational function \( u_P \), zero at \( P \), with the property: if \( r \) is any rational function not identically zero, then

\[
r = u_P^d s,
\]

for some integer \( d \) and some rational function \( s \) that is finite and non-zero at \( P \). Furthermore, the number \( d \) does not depend on the choice of the function \( u_P \). Such a function \( u_P \) is called a uniformizer or uniformizing variable at \( P \).
This theorem allows us to make the following definition.

**Definition 5.3.3** If \( r \) is a rational function and \( r = u_P^d s \), where \( u_P \) is a uniformizer at \( P \), we say that the order of \( r \) at \( P \) is \( d \) and write

\[
\text{ord}_P(r) = d.
\]

If \( r \) has a zero or pole at \( P \), we define its multiplicity to be \( |\text{ord}_P(r)| \).

**Example.** Let \( f \) be a non-zero polynomial on \( E \). We wish to find its order at \( O \). The rational function \( u = x/y \) is a uniformizer at \( O \). To see this, note that \( \deg(x) = 2 < 3 = \deg(y) \), and therefore, \( u(O) = 0 \). Furthermore,

\[ f = \left( \frac{x}{y} \right)^d \left( \frac{y^d f}{x^d} \right), \]

for any integer \( d \). The order of \( f \) is the integer \( d \) that makes \( y^d f/x^d \) finite and non-zero at \( O \). From Definition 5.2.6 this means

\[
\deg(y^d f) = \deg(x^d) \iff 3d + \deg(f) = 2d \iff \deg(f) = -d.
\]

In other words, for every non-zero polynomial \( f \) on \( E \),

\[
\text{ord}_O(f) = -\deg(f). \quad (5.2)
\]

\( \Box \)

Let \( r = f/g \) be a rational function. If we apply Theorem 5.3.2 to the polynomials \( f \) and \( g \) at a point \( P \in E \) we get

\[ f = u_P^{d_1} s_1 \quad \text{and} \quad g = u_P^{d_2} s_2 \]

for rational functions \( s_1 \) and \( s_2 \) finite and non-zero at \( P \). Then we obtain

\[ r = \frac{f}{g} = u_P^{d_1 - d_2} \frac{s_1}{s_2} \]

and \( s_1/s_2 \) is finite and non-zero at \( P \). Thus, the order of \( r \) at \( P \) is \( d_1 - d_2 \). We have established

\[
\text{ord}_P(r) = \text{ord}_P(f) - \text{ord}_P(g).
\]

The next lemma gives a connection between the degree of a polynomial and its zeros and poles. The proof can be found in [12, p.9].
Lemma 5.3.4 Let \( f \) be a polynomial on \( E \). The sum of the multiplicities of the zeros of \( f \) equals the degree of \( f \).

This lemma, together with Equation (5.2), implies that
\[
\sum_{P \in E} \text{ord}_P(f) = \sum_{P \in E - O} \text{ord}_P(f) + \text{ord}_O(f) = 0.
\]

Since for every rational function \( r = f/g \), \( \text{ord}_P(r) = \text{ord}_P(f) - \text{ord}_P(g) \) we have the following important theorem.

Theorem 5.3.5 Let \( r \) be a non-zero rational function on \( E \). Then
\[
\sum_{P \in E} \text{ord}_P(r) = 0.
\]

### 5.4 Divisors

Let \( E \) be an elliptic curve over \( K \). A divisor of \( E \) is a formal sum of the form
\[
\sum_{P \in E} n_P(P),
\]
where \( n_P \in \mathbb{Z} \), and only finitely many of those are non-zero. We call the set of points \( P \) for which \( n_P \neq 0 \) the support of the divisor. Then the set of divisors of \( E \) form a group under the addition law:
\[
\sum_{P \in E} n_P(P) + \sum_{P \in E} m_P(P) = \sum_{P \in E} (n_P + m_P)(P).
\]

This is called the divisor group of \( E \), and is denoted by \( \text{Div}(E) \). The degree of a divisor \( D \in \text{Div}(E) \), is defined by
\[
\deg(D) = \sum_{P \in E} n_P.
\]

The divisors of degree 0 form a subgroup of \( \text{Div}(E) \), which is denoted by \( \text{Div}^0(E) \). With every rational function \( r \) on \( E \) we can associate the following divisor
\[
\text{div}(r) = \sum_{P \in E} \text{ord}_P(r)(P).
\]

Theorem 5.3.5 then implies that for every non-zero rational function \( r \) on \( E \), \( \text{div}(r) \in \text{Div}^0(E) \). Thus, the divisor of any rational function has degree zero. The converse is not true.
**Definition 5.4.1** A divisor $D \in \text{Div}(E)$ is principal if it is the divisor of some function $r$, that is $D = \text{div}(r)$. The group of principal divisors is denoted by $\text{Prin}(E)$. Two divisors $D_1, D_2$ are linearly equivalent, denoted $D_1 \sim D_2$, if $D_1 - D_2$ is principal. The divisor class group (or Picard group) of $E$, denoted $\text{Pic}(E)$ is defined by

$$\text{Pic}(E) = \frac{\text{Div}(E)}{\text{Prin}(E)}.$$ 

If we restrict our attention to divisors of degree 0, then we have the following definition.

**Definition 5.4.2** The degree zero part of the divisor class group of $E$, denoted by $\text{Pic}^0(E)$, is the quotient

$$\text{Pic}^0(E) = \frac{\text{Div}^0(E)}{\text{Prin}(E)}.$$ 

The group $\text{Pic}^0(E)$ is also called the Jacobian of the curve $E$. For elliptic curves, it turns out that the group $\text{Pic}^0(E)$ is in one-to-one correspondence with the set $E(K)$. This correspondence induces an addition law on the set of points of $E$, which makes $E(K)$ a group.

### 5.5 The group law

Let $E$ be an elliptic curve. Consider two points $P$ and $Q$ in $E$, and the line that contains them. Since the Weierstrass equation has degree 3, there is always a third point of intersection $R$ with the curve ($P, Q$ and $R$ may not be distinct, for instance if the line is tangent to $E$). The composition law $\oplus$ is given by the following rule.

**Definition 5.5.1 (Composition Law)** Let $P, Q \in E$, $L$ the line connecting $P$ and $Q$ (tangent line to $E$ if $P = Q$), and $R$ the third point of intersection of $L$ with $E$. Let $L'$ be the line connecting $R$ and $O$ (the vertical line through $R$). Then $P \oplus Q$ is defined to be the third point of intersection of $L'$ with $E$.

It is clear from the definition of the composition law, that the “sum” of two points $P$ and $Q$ can be given by explicit algebraic formulas. In fact, $P \oplus Q$ can be computed efficiently. Furthermore, one can show that the composition is associative, commutative, the point at infinity $O$ serves as the identity, and every point has an inverse. We will not give the formulas here, since they will not be used in the rest of the thesis. They can be found in [64, pp. 58-59].

The above geometric composition law (which seems artificial) can be derived in an indirect way, namely by establishing the isomorphism between $E$ and $\text{Pic}^0(E)$ (as sets). Those facts are recorded in the following theorem.
Theorem 5.5.2 Let $E$ be an elliptic curve.

1. For every $P, Q \in E$, $(P) \sim (Q) \iff P = Q$.

2. For every divisor $D \in \text{Div}^0(E)$ there exists a unique point $P \in E$ such that
   
   $D \sim (P) - (O)$.

3. Let $\sigma : \text{Div}^0(E) \to E$ be the map given by the above association. Then $\sigma$ is surjective.

4. Let $D_1, D_2 \in \text{Div}^0(E)$. Then
   
   $\sigma(D_1) = \sigma(D_2) \iff D_1 \sim D_2$.

   This induces a bijection of sets (which we also denote by $\sigma$)
   
   $\sigma : \text{Pic}^0(E) \tilde{\to} E$.

5. The inverse to $\sigma$ is the map
   
   $\kappa : E \tilde{\to} \text{Pic}^0(E)$
   
   $P \mapsto \text{class of } (P) - (O)$

6. The group law given in Definition 5.5.1 and the group law induced from $\text{Pic}^0(E)$ by $\sigma$ are the same.

As a consequence of Theorem 5.5.2, $E$ becomes an abelian group. From here on, we will drop the special symbol $\oplus$ and simply use $+$ for the group operation on $E$. We denote the inverse of $P$ by $-P$.

We introduce the following notation: for $m \in \mathbb{Z}$ and $P \in E$, we write

\begin{align*}
[0]P &= O, \\
[m]P &= P + P + \cdots + P \text{ (m terms)} \quad \text{for } m > 0, \\
[m]P &= [-m](-P) \quad \text{for } m < 0.
\end{align*}

The following corollary of Theorem 5.5.2 gives a necessary and sufficient condition for a divisor to be principal.

Corollary 5.5.3 Let $E$ be an elliptic curve and $D = \sum n_P(P) \in \text{Div}(E)$. Then $D$ is principal if and only if $\sum n_P = 0$ and $\sum [n_P]P = O$. (Note that the first sum is of integers, the second is addition on $E$.)

More generally, for a (not necessarily principal) divisor $D = \sum n_P(P)$ we say that $D$ sums to $Q$ if $\sum [n_P]P = Q$. 
5.6 Lines

In Theorem 5.5.2 of the previous section we defined the homomorphism

\[ \sigma : \text{Div}^0(E) \to E, \]

that maps \((P) - (O)\) to \(P\). Then, we have

\[ \sigma ((P) + (Q) - 2(O)) = \sigma ((P) - (O)) + \sigma ((Q) - (O)) = P + Q = \sigma ((P + Q) - (O)). \]

Therefore,

\[ (P) + (Q) - 2(O) \sim (P + Q) - (O), \]

which implies that

\[ (P) + (Q) - 2(O) = (P + Q) - (O) + \text{div}(r) \quad \text{for some } r \in \overline{K}(E). \]

Similarly we have

\[ (P) - (Q) = (P - Q) - (O) + \text{div}(r) \quad \text{for some } r \in \overline{K}(E). \]

Thus, for any divisor \(D = \sum n_P(P) \in \text{Div}^0(E)\) we can write

\[ D = \sum n_P ((P) - (O)), \]

since \(\sum n_P = 0\), and by the previous argument, we have

\[ D = \left(\sum [n_P]P\right) - (O) + \text{div}(r) \quad \text{for some } r \in \overline{K}(E). \]

The goal of this section is to show that the function \(r\) can be computed efficiently. For that we need to introduce the notion of a line.

A line is a polynomial on \(E\) of the form

\[ \ell(x, y) = \alpha x + \beta y + \gamma, \quad \alpha, \beta, \gamma \in \overline{K} \]

where at least one of \(\alpha, \beta\) is non-zero. Then one can prove that any line has one of the following divisors.

1. \(\text{div}(\ell) = (P) + (Q) + (R) - 3(O), \quad P, Q, R \text{ distinct,}\)
2. \(\text{div}(\ell) = 2(P) + (Q) - 3(O), \quad \text{tangent line at } P,\)
3. \(\text{div}(\ell) = 3(P) - 3(O), \quad P \text{ is an inflection point,}\)
4. \( \text{div}(\ell) = (P) + (Q) - 2(O) \), \( P, Q = -P \),

5. \( \text{div}(\ell) = 2(P) - 2(O) \), \( \text{vertical tangent line at} P \),

for distinct points \( P, Q \in E \). Conversely, each of the above divisors is the divisor of a line.

Lines are important when we want to solve the following problem. Given a divisor \( D \in \text{Div}^0(E) \), find a point \( R \) and a rational function \( r \in \mathcal{K}(E) \) such that \( D = (R) - (O) + \text{div}(r) \). Let \( D_1 = (P) - (O) \) and \( D_2 = (Q) - (O) \). Suppose that \( D = D_1 + D_2 = (P) + (Q) - 2(O) \), and \( P = (a, b) \) and \( Q = (c, d) \). We need to consider two cases:

**If** \( Q = -P \). Then \( D = (P) + (Q) - 2(O) \) and therefore, \( D = \text{div}(\ell) \), where \( \ell(x, y) = x - a \) is the vertical line through \( P \) (case 4).

**If** \( Q \neq -P \). Then the line that intersects \( E \) at \( P \) and \( Q \), has a third point of intersection \( P \ast Q \) with the curve. We consider the lines \( \ell \) and \( h \) with divisors

\[
\text{div}(\ell) = (P) + (Q) + (P \ast Q) - 3(O), \\
\text{div}(h) = (P \ast Q) + (P + Q) - 2(O).
\]

Then we have

\[
D - \text{div}(\ell) + \text{div}(h) = (P + Q) - (O) \quad \Rightarrow \quad D = (P + Q) - (O) + \text{div}(\ell/h).
\]

where \( \ell(x, y) = y - ax - \gamma \) is the line that contains both \( P \) and \( Q \), and \( h(x, y) = x - x(P + Q) \) is the vertical line through \( P + Q \) and \( P \ast Q \).

Clearly, if we have a finite number of divisors \( D_i = (P_i) - (O) \), \( i = 1, \ldots, n \), we can repeat the process to end up with a rational function \( r \in \mathcal{K}(E) \) such that

\[
D_1 + \cdots + D_n = (P_1 + \cdots + P_n) - (O) + \text{div}(r). \quad (5.5)
\]

We consider now the case \( D = D_1 - D_2 = (P) - (Q) \).

**If** \( Q = -P \). Then \( D = (P) - (Q) \) and if \( h \) is the vertical line through \( P \) and \( Q \) (case 4), we have

\[
\text{div}(h) = (P) + (Q) - 2(O),
\]

and therefore,

\[
D + \text{div}(h) = 2(P) - 2(O) \quad \Rightarrow \quad D = \text{div}(1/h) + 2(P) - 2(O),
\]
and from Equation (5.5), we have
\[ 2(P) - 2(O) = (2P) - (O) + \text{div}(r), \]
for a function \( r \) that is computable. Thus,
\[ D = (2P) - (O) + \text{div}(r/h). \]

If \( Q \neq -P \). Then \( D = (P) - (Q) \) and \( \ell \) is the line that contains \( P, Q \) and \( P \ast Q \), then
\[ D + \text{div}(\ell) = 2(P) + (P \ast Q) - 3(O). \]

Again from Equation (5.5), we can compute a rational function \( r \) such that
\[ 2(P) + (P \ast Q) - 3(O) = (2P + P \ast Q) - (O) + \text{div}(r) = (P - Q) - (O) + \text{div}(r). \]
Therefore,
\[ D = (P - Q) - (O)\text{div}(r/\ell). \]

We observe that if \( D_1 = (P) - (O) \) and \( D_2 = (Q) - (O) \) then
\[ D_1 \pm D_2 = (P \pm Q) - (O) + \text{div}(r), \]
and the function \( r \) is efficiently computable.

We can use the above idea to compute a rational function \( r \in \overline{K}(E) \) such that \( D = \text{div}(r) \), for any divisor \( D \in \text{Prin}(E) \). Indeed, if
\[ D = \sum_{i=1}^{m} n_i(P_i), \]
then we can write
\[ D = \sum_{i=1}^{m} n_i((P_i) - (O)), \]
since \( \sum n_i = 0 \) by the principality of \( D \). Then, using the above algorithm, we find a function \( r \in \overline{K}(E) \) such that
\[ D = (\sum_{i=1}^{m}[n_i]P_i) - (O) + \text{div}(r). \]
Since \( D \) is principal, we have
\[ \sum_{i=1}^{m}[n_i]P_i = O. \]
Therefore,
\[ D = \text{div}(r). \]
5.7 Maps between elliptic curves

We define now the notion of a \textit{rational map} from an elliptic curve \( E \) to itself. We note that everything in this section can be generalized to rational maps between two elliptic curves \( E \) and \( E' \). Let \( f(X,Y) \) be the defining polynomial for \( E \).

\textbf{Definition 5.7.1} A \textit{rational map} \( F \) from \( E \) to \( E \) is a pair \( (r,s) \), where \( r, s \) are rational functions on \( E \) such that

\[ f(r,s) = 0. \]

If we let \( F(P) = O \) if and only if \( r \) and \( s \) are not finite at \( P \), then \( F \) defines a map

\[
F : E \rightarrow E, \\
P \mapsto (r(P), s(P)),
\]

since \( r \) and \( s \) have poles at the same points, and if they are finite at \( P \) then \( (r(P), s(P)) \) is a point on \( E \).

\textbf{Definition 5.7.2} Let \( F \) be a \textit{rational map} from \( E \) to itself. The \textit{ramification index} of \( F \) at \( P \in E \) is defined by

\[ e_F(P) = \text{ord}_P(u_{F(P)} \circ F), \]

where \( u_{F(P)} \) is a uniformizer at \( F(P) \in E \).

To a rational map \( F \) corresponds in a natural way a map

\[
F^* : K(E) \rightarrow K'(E) \\
r \mapsto r \circ F
\]

It is not hard to see that \( F^* \) is a field homomorphism.

\textbf{Theorem 5.7.3} Every non-constant rational map \( F : E \rightarrow E \) is onto.

Since \( F \) is onto, it follows that \( F^* \) is one-to-one. Indeed, if \( F^*(r) = F^*(s) \) then for every \( P \in E \) we have \( r(F(P)) = s(F(P)) \). From the surjectivity of \( F \), it follows that \( F(P) \) varies over all points of \( E \). Thus, \( r(Q) = s(Q) \) for all \( Q \in E \).

\textbf{Theorem 5.7.4} Let \( F : E \rightarrow E \) be a non-constant rational map. Then \( F^* : K(E) \rightarrow K'(E) \) is a field homomorphism.
It is clear that $F^*K(E)$ is a subfield of $K(E)$. In fact, it can be proven that $K(E)/F^*K(E)$ is a finite field extension. We define the degree of this extension to be the degree of $F$, i.e.,

$$\deg(F) = [K(E) : F^*K(E)].$$

We define the separable (resp. inseparable) degree of $F$, denoted $\deg_s(F)$ (resp. $\deg_i(F)$) to be the separable (resp. inseparable) degree of the field extension. The usefulness of this definition will become clear in Section 5.8.

A rational map $F$ of $E$ induces a map of $\text{Div}(E)$, which is usually also denoted by $F^*$, given by

$$F^* : \text{Div}(E) \to \text{Div}(E) \quad (Q) \mapsto \sum_{P \in F^{-1}(Q)} e_F(P)(P).$$

The map is extended $\mathbb{Z}$-linearly to any divisor. The next theorem gives some properties of the map $F^*$, and shows the connection between the degrees of rational maps and the degrees of divisors. We stress that $F^*$ denotes two different maps (one on $\overline{K}(E)$ and one on $\text{Div}(E)$). Which one is meant each time will be clear from the argument of $F^*$.

**Theorem 5.7.5** Let $F : E \to E$ be a non-constant map. Then,

1. For every $r \in \overline{K}(E)^*$, $\text{ord}_P(F^*r) = e_F(P)\text{ord}_{F(P)}(r)$.

2. For every $D \in \text{Div}(E)$, $\deg(F^*D) = \deg(F)\deg(D)$.

3. For every $r \in \overline{K}(E)^*$, $F^*(\text{div}(r)) = \text{div}(F^*r)$.

We will be interested in a special class of rational maps from $E$ to $E$, namely those that are also group homomorphisms.

**Definition 5.7.6** A rational map from $E$ to $E$ that is also a group homomorphism is called an endomorphism. These maps form a ring, which is denoted by $\text{End}(E)$. The operations are defined by

$$F + G)(P) = F(P) + G(P) \quad \forall P \in E,$$

$$F \circ G)(P) = F(G(P)) \quad \forall P \in E.$$

The endomorphism ring is, in general, not commutative, i.e., the operator $\circ$ does not commute. Some useful properties of $\text{End}(E)$ are given in the next proposition.
Proposition 5.7.7 Let $E$ be an elliptic curve. Then the endomorphism ring $\text{End}(E)$ is a (not necessarily commutative) ring of characteristic 0 with no zero divisors.

Example.

1. The “multiplication by $m$” map $[m] : E \to E$, defined by Equation (5.4) is an endomorphism.

2. Let $K = \mathbb{F}_q$, $q$ a prime power, and $E$ an elliptic curve defined over $\mathbb{F}_q$. The Frobenius map $\phi : E \to E$ is given by $\phi = (x^q, y^q)$. In other words,

$$
\phi(O) = O, \quad \phi(P) = (a^q, b^q), \text{ for } P = (a, b) \in E - O.
$$

The Frobenius map is an endomorphism of $E$. To see this, one needs to notice that for every $a \in \mathbb{F}_q$, $a^q = a$, and the addition on $E$ is given by rational functions (with coefficients in $\mathbb{F}_q$). We stress that $\phi$ is not the identity map. However, it is true that $\phi(P) = P$ for every $P \in E(\mathbb{F}_q)$ (while this is not true for points defined over extensions of $\mathbb{F}_q$). In fact, the property $\phi(P) = P$ characterizes the points on $E$ defined over $\mathbb{F}_q$. This is the reason that the Frobenius map plays an important role in the theory of elliptic curves over finite fields. We will use the Frobenius map extensively in Chapter 6.

3. The “translation by $T$” map $\tau_T : E \to E$ defined by $\tau_T(P) = P + T$ is a rational map, which is not an endomorphism, unless $T = O$.

\[ \square \]

5.8 Isogenies

We introduce now a seemingly more general class of rational maps, called isogenies. Again, isogenies are more generally defined as maps between two elliptic curves $E$ and $E'$ with some specific property. We will be interested only in maps from $E$ to itself.

Definition 5.8.1 An isogeny is a rational map $F : E \to E$ with the property $F(O) = O$.

Clearly every endomorphism is an isogeny, since an endomorphism is a homomorphism of groups, and therefore maps the identity to the identity. The converse turns out to be also true.
Theorem 5.8.2 Every isogeny $F : E \to E$ is also an endomorphism.

Nonzero isogenies have several special properties. For instance, the ramification index $e_F(P)$ does not depend on the point $P$.

Theorem 5.8.3 Let $F : E \to E$ be a non-constant isogeny.

1. For every $P \in E$, $\# F^{-1}(P) = \deg_e(F)$.

2. For every $P \in E$, $e_F(P) = \deg_i(F)$.

3. The map

$$
\ker(F) \to \text{Aut}[\overline{K}(E)/F^*\overline{K}(E)]
$$

$$
T \mapsto \tau_T^*
$$

is an isomorphism. Here $\tau_T$ is the translation by $T$ map and $\tau_T^*$ is the automorphism it induces on $\overline{K}(E)$.

4. Assume that $F$ is separable. Then $F$ is unramified, and $\# \ker(F) = \deg(F)$, and $\overline{K}(E)$ is Galois over $F^*\overline{K}(E)$.

This theorem gives a more intuitive understanding of the notions of degree, separable degree and inseparable degree, for non-constant isogenies.

We apply now Theorem 5.8.3 to the $q$-power Frobenius endomorphism of $E/\mathbb{F}_q$. We know that $e_\phi$ does not depend on the point, so we can write

$$
e_\phi = e_\phi(O) = \text{ord}_O(u \circ \phi),
$$

where the second equality is by the definition of the ramification index and $u = x/y$ is a uniformizer at $O$ (as we saw in Section 5.3). Since $u \circ \phi = u^q$, we have

$$
e_\phi = \text{ord}_O(u^q) = q.
$$

Furthermore, we can see that $\phi^* \mathbb{F}_q(E)$ is the subfield given by quotients

$$
\phi^* \left( \frac{f}{g} \right) = \frac{f(x^q, y^q)}{g(x^q, y^q)}.
$$

Similarly, $\mathbb{F}_q(E)^q$ is the subfield given by quotients

$$
\frac{f(x, y)^q}{g(x, y)^q}.
$$
We also know that 

$$(\mathbb{F}_q[x,y])^q = \mathbb{F}_q[x^q, y^q].$$

Therefore, we have $\phi^* \mathbb{F}_q(E) = \mathbb{F}_q(E)^q$. This implies that the extension $\mathbb{F}_q(E)/\phi^* \mathbb{F}_q(E)$ is purely inseparable, because for every element $r \in \mathbb{F}_q(E)$, $r^q \in \phi^* \mathbb{F}_q(E)$ (for more details see [46, p.45]). These observations combined with Theorem 5.8.3 imply the following theorem.

**Theorem 5.8.4** Let $E/\mathbb{F}_q$ be an elliptic curve, and $\phi$ the $q$-power Frobenius endomorphism. Then

1. $e_\phi = \deg_q(\phi) = q$.
2. $\phi$ is purely inseparable.
3. $\deg_q(\phi) = 1$, which implies $\ker(\phi) = \{O\}$. 

The “multiplication-by-$m$” map, on the other hand, is separable, provided that the characteristic of the field does not divide $m$.

**Theorem 5.8.5** Let $E/K$ be an elliptic curve, and let $m \in \mathbb{Z}$, $m \neq 0$. If $\text{char}(K) = 0$ or $\text{char}(K)$ is prime to $m$ then the endomorphism $[m] : E \to E$ is separable.

This, together with Theorem 5.8.3, imply that for $E/\mathbb{F}_q$ with $m$ prime to $q$,

$$\# \ker[m] = \deg[m]. \quad \text{and} \quad e_m = 1.$$ 

A similar result holds for endomorphisms of the form $[m] + [n] \circ \phi$, which we denote by $m + n\phi$.

**Theorem 5.8.6** Let $E/\mathbb{F}_q$ be an elliptic curve, $\phi$ the $q$-power Frobenius endomorphism, and $m, n \in \mathbb{Z}$. Then the map

$$m + n\phi : E \to E$$

is separable if and only if $\gcd(m, q) = 1$. In particular, the endomorphism $1 - \phi$ is separable.

The endomorphism $1 - \phi$ is particularly important. For a point $P \in E(\mathbb{F}_q)$

$$P \in E(\mathbb{F}_q) \iff \phi(P) = P \iff P \in \ker(1 - \phi).$$

Thus the order of the group $E(\mathbb{F}_q)$ is $\# \ker(1 - \phi)$, and since $1 - \phi$ is separable, it is equal to $\deg(1 - \phi)$. Hasse in the 1930's gave an estimate of $\deg(1 - \phi)$ proving the following famous theorem.
Theorem 5.8.7 (Hasse) Let $E/\mathbb{F}_q$ be an elliptic curve. Then
\[
\#E(\mathbb{F}_q) = q + 1 - a_q, \quad \text{with } |a_q| \leq 2\sqrt{q}.
\]
A proof of Theorem 5.8.7 can be found in [64, pp. 131-132]. The number $a_q$ is sometimes called the trace of Frobenius.

5.9 The dual isogeny

To every non-constant isogeny $F$ corresponds a unique isogeny, called the dual isogeny, denoted by $\hat{F}$. This correspondence is given by the following theorem.

Theorem 5.9.1 Let $F : E \to E$ be a non-constant isogeny of degree $m$. There exists a unique isogeny $\hat{F} : E \to E$ satisfying
\[
\hat{F} \circ F = [m].
\]

Definition 5.9.2 Let $F : E \to E$ be an isogeny. The dual isogeny to $F$ is the isogeny given by Theorem 5.9.1. If $F$ is constant (i.e., $F = [0]$), then we set $\hat{F} = [0]$.

The next theorem gives the basic properties of the dual isogeny, that will be used extensively in Chapter 6.

Theorem 5.9.3 Let $F : E \to E$ be an isogeny.

1. Let $m = \deg(F)$. Then
\[
\hat{F} \circ F = F \circ \hat{F} = [m].
\]

2. Let $G : E \to E$ be another isogeny. Then
\[
\hat{G} \circ \hat{F} = \hat{F} \circ \hat{G},
\]
\[
\hat{G} + \hat{F} = \hat{G} + \hat{F}.
\]

3. For all $m \in \mathbb{Z}$, $[\hat{m}] = [m]$ and $\deg([m]) = m^2$.

4. $\deg(\hat{F}) = \deg(F)$.

5. $\hat{\hat{F}} = F$.

From Theorem 5.9.3 and the separability of the map $[m]$ (Theorem 5.8.5), we get the following theorem.

Theorem 5.9.4 Let $E/K$ be an elliptic curve and $m \in \mathbb{Z}$, $m \neq 0$. 

1. If \( \text{char}(K) \) is zero or prime to \( m \), then
\[
E[m] \cong \frac{\mathbb{Z}}{m\mathbb{Z}} \oplus \frac{\mathbb{Z}}{m\mathbb{Z}}.
\]

2. If \( \text{char}(K) = p \), then either
\[
E[p^k] = \{O\} \quad \text{for all } k = 1, 2, \ldots; \quad \text{or}
\]
\[
E[p^k] = \frac{\mathbb{Z}}{p^k\mathbb{Z}} \quad \text{for all } k = 1, 2, \ldots.
\]

(Notation: \( E[m] \) is another notation for \( \ker[m] \), the set of points of \( E \) having order that divides \( m \).)

Let \( \text{char}(K) = p \). The elliptic curves \( E/K \) for which \( E[p^k] = \{O\} \) are called supersingular. A curve that is not supersingular is called ordinary.

Let \( E/F_q \) be an elliptic curve. We know that \( E(F_q) \) form a group, necessarily finite, since \( F_q \) is finite. Moreover, Hasse's Theorem 5.8.7 gives an estimate of the order of the group. As a matter of fact, more can be shown about the structure of \( E(F_q) \). The following result is easily deduced from Theorem 5.9.4.

**Theorem 5.9.5** Let \( E \) be an elliptic curve over \( F_q \). Then
\[
E(F_q) \cong \frac{\mathbb{Z}}{n_1\mathbb{Z}} \oplus \frac{\mathbb{Z}}{n_2\mathbb{Z}},
\]
for some positive integers \( n_1, n_2 \), with \( n_2|n_1 \) and \( n_2|q - 1 \).

It is immediate from the theorem that there exist points \( T_1 \) and \( T_2 \) that generate the group \( E(F_q) \), i.e., for every point \( P \in E(F_q) \), \( P = [x]T_1 + [y]T_2 \), with \( x \in \mathbb{Z}/n_1\mathbb{Z} \), and \( y \in \mathbb{Z}/n_2\mathbb{Z} \). In the case that \( n_2 = 1 \) the group is cyclic.

### 5.10 Weil reciprocity

In this section, we describe the Weil reciprocity, and the generalized Weil reciprocity. The later will be useful in Chapter 6, for the computation of the Weil pairing.

We start by defining the value of a function at a divisor.

**Definition 5.10.1** Let \( E \) be an elliptic curve, and \( f \in \overline{K}(E)^* \) a non-zero rational function. Let \( D = \sum n_P(P) \) be a divisor such that \( D \) and \( \text{div}(f) \) have disjoint support. Then we define the value of \( f \) at \( D \) as
\[
f(D) = \prod_{P \in E} f(P)^{n_P}.
\]
The Weil reciprocity law can be stated as follows.

**Theorem 5.10.2 (Weil reciprocity)** Let $E$ be an elliptic curve, and $f, g \in \overline{K}(E)^\ast$, such that $\text{div}(f)$ and $\text{div}(g)$ have disjoint support. Then

$$f(\text{div}(g)) = g(\text{div}(f)).$$

There is a generalization of the Weil reciprocity law, that removes the requirement that $\text{div}(f)$ and $\text{div}(g)$ have disjoint support. Our presentation is from [12]. We start by defining the *local symbol* of $f$ and $g$.

**Definition 5.10.3** Let $E$ be an elliptic curve and $f, g \in \overline{K}(E)^\ast$. Assume that $m_f = \text{ord}_P(f)$ and $m_g = \text{ord}_P(g)$. Then we define

$$\langle f, g \rangle_P = (-1)^{m_fm_g} \frac{f^{m_g}}{g^{m_f}}(P).$$

We note, that the local symbol is well defined, since the function $f^{m_g}/g^{m_f}$ has order zero at $P$. Indeed,

$$\text{ord}_P \left( \frac{f^{m_g}}{g^{m_f}} \right) = \text{ord}_P(f^{m_g}) - \text{ord}_P(g^{m_f})$$

$$= m_g \text{ord}_P(f) - m_f \text{ord}_P(g)$$

$$= 0$$

We also note that the local symbol is that $\langle f, g \rangle_P = 1$ unless $f$ or $g$ has a zero or a pole at $P$. This is simply because if both $f$ and $g$ are finite and non-zero at $P$ then $m_f = m_g = 0$. The generalized Weil reciprocity law is given by the next theorem.

**Theorem 5.10.4 (Generalized Weil reciprocity)** Let $E$ be an elliptic curve, and $f, g \in \overline{K}(E)^\ast$. Then

$$\prod_{P \in E} \langle f, g \rangle_P = 1.$$
Chapter 6

The elliptic curve discrete logarithm problem

6.1 Introduction

The first public-key cryptographic protocol, that was proposed by Diffie and Hellman [17] in 1976 based its security on the presumed intractability of the discrete logarithm problem in the multiplicative group of finite fields. The intense research during the following years resulted in the Index Calculus method and its several variants — for instance see Chapter 4 — that run in subexponential time. The most successful such variant is the Number Field Sieve [55, 56, 57], that has conjectured running time

$$\exp \left( (c + o(1))(\log q)^{1/3}(\log \log q)^{2/3} \right).$$

This development made the DLP in finite fields tractable for finite fields $\mathbb{F}_q$ with $q \sim 2^{1000}$. Consequently, the security of cryptosystems based on the DLP in finite fields is compromised, unless quite large keys are used.

The above developments motivated researchers to look for groups other than the multiplicative group of finite fields, in which the DLP remained of exponential time complexity, and that could be used as the basis of a cryptographic system. One such proposal is to use the group of points of an elliptic curve over a finite field.

Elliptic curves have features that are attractive for cryptographic purposes. To start with, the arithmetic in $E(\mathbb{F}_q)$ is efficient (see for instance [47, 67]). Furthermore, Hasse’s theorem asserts that the order of the group $#E(\mathbb{F}_q)$ is $q + 1 - a_q$ with $|a_q| \leq 2\sqrt{q}$, i.e., very close to the size of $\mathbb{F}_q$ itself. It is also important that $#E(\mathbb{F}_q)$, which is needed for some cryptosystems, can be computed efficiently using the algorithms of Schoof [58], Lehmann
et.al. [35], and Menezes et.al. [43]. Groups which are candidates for cryptographic use, with smooth order can also be avoided this way. In addition, elliptic curves define a great variety of finite abelian groups, that are candidate for cryptographic use.

Probably the most important aspect of a cryptosystem is the level of security that it offers. In our case, this depends on the difficulty of the discrete logarithm problem on elliptic curves (ECDLP), which is the subject of this chapter.

6.2 Reductions to finite fields

As mentioned before, the ECDLP remains, in general, of exponential time complexity; that is, no subexponential time algorithm is known that works for every elliptic curve $E$ over any finite field $\mathbb{F}_q$. One reason for that is that the Index Calculus type methods do not seem to apply [65]. However, not all elliptic curves offer the same level of security. There exist elliptic curves, for which subexponential algorithms have been found. The method of attack has, invariably, been the construction of a polynomial time reduction of the ECDLP to DLP in either the additive or the multiplicative group of finite fields. Then, a subexponential algorithm for the latter problem can be employed.

Let us fix some notation here. We consider an elliptic curve $E$ defined over a finite field $\mathbb{F}_q$ of characteristic $p$. We are given a point $P \in E(\mathbb{F}_q)$ of prime order $r \neq p$, and a point $Q \in \langle P \rangle$ whose discrete logarithm we want to compute. All the attacks that we discuss in this thesis use an efficiently computable isomorphism

$$\langle P \rangle \xrightarrow{\sim} \mu_r,$$

where $\mu_r$ is the group of $r$-th roots of unity in $\mathbb{F}_q$. The group $\mu_r$ is contained in the extension field $\mathbb{F}_{q^k}$, where $k$ is the smallest positive integer such that $r | q^k - 1$. Therefore, such an isomorphism can be effective only if the extension degree $k$ is small. We will consider extensively the case $r | q - 1$, i.e., all the $r$-th roots of unity are contained in $\mathbb{F}_q$. The goal of this chapter is to prove the following theorem.

**Theorem 6.2.1** Let $E/\mathbb{F}_q$ be an elliptic curve, $P \in E(\mathbb{F}_q)$ a point of prime order $r$ different than the characteristic of $\mathbb{F}_q$, and assume that $a_q \equiv 2 \pmod{r}$. Let $\mu_r$ denote the group of $r$-th roots of unity in $\mathbb{F}_q$. Then there exists an isomorphism

$$\langle P \rangle \xrightarrow{\sim} \mu_r,$$

that is computable in polynomial time.
The proof is broken down into two parts. First we construct the isomorphism using a generalized version of the Weil pairing. Then we give an algorithm to compute the generalized Weil pairing in the case of interest to us. At the end of the chapter we will make some remarks about the more general case, \( r | q^k - 1 \) for \( k > 1 \).

Before moving to our contribution, we review the two main results in the literature.

### 6.2.1 The MOV reduction

Let \( E / \mathbb{F}_q \) be an elliptic curve defined over \( \mathbb{F}_q \), and let \( P \in E(\mathbb{F}_q) \) be a point of prime order \( r \), different from the characteristic of \( \mathbb{F}_q \). We wish to solve the discrete logarithm problem in \( \langle P \rangle \); that is, given a point \( Q \in \langle P \rangle \) we want to find an integer \( \ell \) such that \( Q = [\ell]P \).

The idea of Menezes, Okamoto and Vanstone [42] is to construct an isomorphism between \( \langle P \rangle \) and the group of \( r \)-th roots of unity in \( \mathbb{F}_q \), which we denote by \( \mu_r \). The construction of the isomorphism uses the Weil pairing \( w \), which we describe now.

Let \( E[r] \) be the group of points of order dividing \( r \). We consider a point \( S \in E[r] \) and a function with divisor

\[
\text{div}(f_S) = r(S) - r(O).
\]

Then we observe that

\[
[r]^* ((S) - (O)) = \sum_{R \in E[r]} (S' + R) - (R),
\]

is a principal divisor, where the point \( S' \) is chosen so that \( [r]S' = S \). We let

\[
\text{div}(g_S) = \sum_{R \in E[r]} (S' + R) - (R).
\]

Then we have

\[
\text{div}(f_S \circ [r]) = r \text{div}(g_S),
\]

and therefore

\[
g_S' = f_S \circ [r].
\]

Of course, the functions \( f_S \) and \( g_S \) are only defined up to a multiplicative constant. The constant can be chosen so that the above equation holds. The Weil pairing is defined by

\[
w : E[r] \times E[r] \rightarrow \mu_r
\]

\[
(S, T) \mapsto \frac{g_S(X + T)}{g_S(X)},
\]

where \( X \in E(\mathbb{F}_q) \) is any point such that \( g_S(X) \neq 0 \). It can be proven (see [64]), that the Weil pairing, as given above, is well defined, and has the following properties.
**Identity.** For every $S \in E[r]$, $w(S, S) = 1$.

**Bilinearity.** For every $S, T_1, T_2 \in E[r]$, $w(S, T_1 + T_2) = w(S, T_1)w(S, T_2)$.

For every $S_1, S_2, T \in E[r]$, $w(S_1 + S_2, T) = w(S_1, T)w(S_2, T)$.

**Non-degeneracy.** If $w(S, T) = 1$ for every $T \in E[r]$, then $S = O$.

Furthermore, there exists a point $S \in E[r]$ such that $w(P, S)$ is a primitive $r$-th root of unity. Then, the map

$$V : \langle P \rangle \to \mu_r$$

$$Q \mapsto w(Q, S),$$

is a group isomorphism. The two groups are cyclic of the same order, and $V$ is a group homomorphism, since

$$V(Q_1 + Q_2) = w(Q_1 + Q_2, S) = w(Q_1, S)w(Q_2, S) = V(Q_1, S) \cdot V(Q_2, S).$$

The algorithm then works as follows.

---

**Algorithm: Reduction (Supersingular curves)**

*Input:* A point $P$ of order $r$, and a point $Q \in \langle P \rangle$

*Output:* An integer $\ell$ such that $Q = [\ell]P$

1. Find a point $S \in E[r]$ such that $w(P, S)$ is a primitive $r$-th root of unity.

2. Compute $\alpha = w(P, S)$ and $\beta = w(Q, S)$.

3. Compute the discrete logarithm $\ell$ of $\beta$ to the base $\alpha$.

4. Output $\ell$.

The reason that the algorithm works is that, if $Q = [\ell]P$ then

$$\beta = w(Q, S) = w([\ell]P, S) = w(P, S)^\ell = \alpha^\ell.$$

There are three issues relevant to the efficiency of the above algorithm. The first is that the points involved have to be defined over a small extension of $\mathbb{F}_q$ (otherwise their representation is simply too large to handle), and also the elements $\alpha$ and $\beta$ have to live in a small extension of $\mathbb{F}_q$. The second issue is the determination of the point $S$. The third issue is the efficient computation of the Weil pairing.
Menezes, Okamoto and Vanstone show, that if $E$ is a supersingular curve (see Theorem 5.9.4), then all the requirements can be met. In particular, for supersingular curves

$$E[r] \subseteq E(F_q^*)$$

for an integer $k \leq 6$. This means that all the points involved are defined over a small extension of $F_q$ (extension of degree at most 6). Furthermore, $\mu_r \subseteq F_q^*$ for the same extension degree $k$. Also, the point $S$ can be found probabilistically, by randomly choosing points in $E[r]$. Finally, the issue of the efficient computation of $w$ is resolved by an algorithm of Miller (see [42, Appendix]).

A few remarks about the MOV reduction are in order. A reduction based on the Weil pairing cannot be efficient unless the point $S \in E[r]$ is defined over a small extension of $F_q$, which in turn requires that

$$E[r] \subseteq E(F_q^*) \text{ for a small } k. \quad (6.2)$$

For supersingular curves, this is the case. Harasawa et. al. give in [30] an extension of this approach (using the Weil pairing as well), that works for non-supersingular curves that satisfy Equation (6.2) for some $k$. However, this is as far as the Weil pairing can be "pushed". For curves such that $r | q - 1$ (which, in general, do not satisfy Equation (6.2)), $\mu_r \subseteq F_q$, i.e., the elements of both groups $\langle P \rangle$ and $\mu_r$ are defined over $F_q$. However, the MOV reduction requires for its computation the point $S$, that is, in general, defined over a very large extension of $F_q$, rendering this approach totally inefficient.

### 6.2.2 The FR reduction

A reduction that works when $r | q - 1$ was given by Frey and Rück in [20] (see also [19]). We will refer to it as the "FR reduction". The FR reduction is based on a different pairing, the Tate-Lichtenbaum pairing. The main result of [20], in our notation is the following.

**Theorem 6.2.2 (Frey-Rück)** Let $r$ be a prime number prime to $q$, and let $\mu_r$ be the group of $r$-th roots of unity in $\overline{F}_q$. Assume that $E(F_q)$ contains a point of order $r$. There exists a pairing

$$\phi_r : E(F_q)/rE(F_q) \times E(F_q)[r] \to \mu_r,$$

that is bilinear and non-degenerate. Moreover, it is computable in $O(\log q)$ steps, where one step is equivalent to one addition on $E$. 

Here, \( E(\mathbb{F}_q)[r] \) denotes the group of points defined over \( \mathbb{F}_q \) that have order dividing \( r \). Returning to our problem, we have a point \( P \) of order \( r \), and we want to solve the ECDLP in \( \langle P \rangle \). The case of interest for cryptographic purposes is when \( r \) is large. To make this precise, we assume that \( r^2 \mid \#E(\mathbb{F}_q) \). Then, for elliptic curves with trace of Frobenius \( a_q = 2 \), i.e., \( \#E(\mathbb{F}_q) = q - 1 \) we have \( r \mid q - 1 \) which implies that \( \mu_r \subseteq \mathbb{F}_q \). Frey, Müller, and Rück show that in this case \( \phi_r(P, P) \) is a primitive \( r \)-th root of unity, and the pairing is given by

\[
\phi_r(Q, P) = (f_P(D_Q))^{(q-1)/r},
\]

where \( f_P \) is a function with divisor

\[
\text{div}(f_P) = r(P) - r(O),
\]

and \( D_Q \) is a divisor that sums to \( Q \).

The actual reduction then, is virtually the same as the MOV reduction, with the Weil pairing substituted with the Tate-Lichtenbaum pairing, and the first step eliminated (since the point \( S \) that makes \( \phi_r \) a primitive \( r \)-th root of unity is known to be \( P \)).

<table>
<thead>
<tr>
<th>Reduction</th>
<th>( E/\mathbb{F}_q )</th>
<th>( E/\mathbb{F}_p )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Weil pairing</td>
<td>( p \mid a_q )</td>
<td>( a_p = 0 )</td>
</tr>
<tr>
<td>Tate pairing</td>
<td>( a_q \equiv 2 \pmod{r} )</td>
<td>( a_p = 2 )</td>
</tr>
<tr>
<td>Generalized Weil pairing</td>
<td>( a_q \equiv 2 \pmod{r} )</td>
<td>( a_p = 2 )</td>
</tr>
</tbody>
</table>

Table 6.1: Reductions of the ECDLP to DLP in finite fields

We summarize the cases that the MOV and FR reductions apply in Table 6.2.2. We also include in the table the case when our reduction applies. It is assumed that the characteristic of \( \mathbb{F}_q \) is \( p \). The criteria when the elliptic curve \( E \) is defined over a prime field \( \mathbb{F}_p \) are particularly simple, and are explicitly stated.

### 6.3 The generalized Weil pairing

In this section, we present a generalized form of the Weil pairing as it appears in [64, p.107]. We construct a bilinear non-degenerate pairing for each endomorphism \( \psi \). As for the rest of this chapter, \( p \) is prime, and \( q = p^k \).

Let \( E \) be an elliptic curve over \( \mathbb{F}_q \). Also let \( \psi : E \to E \) be a non-zero endomorphism of \( E \), and denote its dual by \( \hat{\psi} \). From Theorem 5.9.1, the dual exists, is unique, and
is itself an endomorphism. Let $T \in \ker(\hat{\psi})$; such a point certainly exists, since $\hat{\psi}$ is an endomorphism. We denote by $m$ the degree of $\psi$. Then, the divisor $D = m(T) - m(O)$ has degree zero, and

$$[m]T - [m]O = \psi \circ \hat{\psi}(T) = \psi(O) = O,$$

therefore it is principal. Let $f_T \in \overline{\mathbb{F}}_q(E)$ be a function such that

$$\text{div}(f_T) = m(T) - m(O).$$

We consider now the divisor of $f_T \circ \psi$

$$\text{div}(f_T \circ \psi) = \text{div}(\psi^* f_T) = \psi^* \text{div}(f_T) = m(\psi^*(T) - \psi^*(O)),$$

the second equality being true by Theorem 5.7.5, and the third by the definition of $\psi^*$ (Z-linearity). We note that

$$\psi^*(T) - \psi^*(O) = \sum_{\psi P = T} e_{\psi}(P)(P) - \sum_{\psi R = O} e_{\psi}(R)(R)$$

$$= \deg_i \psi \left( \sum_{\psi R = O} (T' + R) - (R) \right),$$

where $\psi T' = T$. Here we used the fact that $\psi$ is an isogeny, and therefore, by Theorem 5.8.3 we know that $e_{\psi}(P)$ does not depend on $P$, and equals to $\deg_i(\psi)$. The last line of the derivation shows that the divisor is principal, since it has degree zero, and it sums to

$$[\deg_i \psi] \sum_{\psi R = O} T' = [\deg \psi] T' = \hat{\psi} \circ \psi(T') = \hat{\psi} T = O.$$

So it must be the divisor of some function $g_T \in \overline{\mathbb{F}}_q(E)$. Thus,

$$(f_T \circ \psi) = m\text{div}(g_T) = \text{div}(g_T^m),$$

which implies that

$$g_T^m = f_T \circ \psi.$$  \hspace{1cm} (6.3)

g_T is defined up to a multiplicative constant of course. Let now $S \in \ker(\psi)$, and $X$ any point of $E(\overline{\mathbb{F}}_q)$.

$$g_T(X + S)^m = f_T(\psi X + \psi S) = f_T(\psi X) = f_T \circ \psi X = g_T(X)^m.$$
We define the pairing
\[ e_\psi : \ker(\psi) \times \ker(\tilde{\psi}) \rightarrow \mu_m \]
as
\[ e_\psi(S, T) = \frac{g_T(X + S)}{g_T(X)}. \]
(6.4)
The above definition does not depend on the choice of \( X \in E(\F_q) \). Indeed, if \( \tau_S \) denotes the translation by \( S \) map
\[ \tau_S : E \rightarrow E \]
\[ X \mapsto X + S \]
then we can write \( e_\psi(S, T) \) as
\[ e_\psi(S, T) = \frac{g_T \circ \tau_S}{g_T}(X), \]
and the function \( g_T \circ \tau_S / g_T \) is constant. Indeed, we need to note that \( \psi = \psi \circ \tau_S \) because \( S \in \ker(\psi) \). Then,
\[ \text{div}(g_T \circ \tau_S) = \tau_S^* \text{div}(g_T) \]
\[ = \tau_S^* \circ \psi^*((T) - (O)) \]
\[ = (\psi \circ \tau_S)^*((T) - (O)) \]
\[ = \psi^*((T) - (O)) \]
\[ = \text{div}(g_T). \]

Next, we prove two important properties of the pairing.

**Non-degeneracy.** Suppose that \( e_\psi(S, T) = 1 \) for every \( S \in \ker(\psi) \). Then from the definition of \( e_\psi \) we have
\[ g_T(X + S) = g_T(X) \text{ for all } S \in \ker(\psi). \]
This means that \( g_T \) is fixed by every translation \( \tau_S \) with \( S \in \ker(\psi) \). From Theorem 5.8.3, we know that \( \ker(\psi) \) is isomorphic to the group of automorphisms of \( \F_q(E) \) that fix \( \F_q(E) \), and the correspondence is
\[ \ker(\psi) \rightarrow \text{Aut}[\F_q(E) / \psi^* \F_q(E)] \]
\[ S \mapsto \tau_S^*. \]
Thus, \( g_T \) is fixed by every element of \( \text{Aut}[\F_q(E) / \psi^* \F_q(E)] \) and therefore has to be in \( \psi^* \F_q(E) \). This implies that
\[ g_T = \psi^* h = h \circ \psi, \]
for some function $h \in \overline{\mathbb{F}}_q(E)$. Then

$$(h \circ \psi)^m = g_T^m = f_T \circ \psi,$$

so $f_T = h^m$, because $\psi$, as an isogeny, is a surjective map $E \to E$. We conclude

$m \text{div}(h) = \text{div}(f_T) = m(T) - m(O),$

which means $\text{div}(f_T) = (T) - (O)$. By the criterion of principality, a principal divisor has to sum to $O$, so we get $T = O$.

**Bilinearity.** We establish now that $e_\psi$ is linear in both arguments (i.e., it is bilinear). Linearity in the first argument is easy.

$$e_\psi(S_1 + S_2, T) = \frac{g_T(X + S_1 + S_2)}{g_T(X + S_2)} \frac{g_T(X + S_2)}{g_T(X)} = e_\psi(S_1, T) e_\psi(S_2, T).$$

Here, we made use of the fact that in the definition of the pairing, $e_\psi(S, T) = g_T(X + S)/g_T(X)$, for any $X \in E$. Above, we chose the points $X$ and $X + S_2$. We consider now $e_\psi(S, T_1 + T_2)$. Let $f_1, f_2, f_3, g_1, g_2, g_3$ be the functions corresponding to the points $T_1, T_2$, and $T_3 = T_1 + T_2$; that is

$$\text{div}(f_1) = m(T_1) - m(O)$$
$$\text{div}(f_2) = m(T_2) - m(O)$$
$$\text{div}(f_3) = m(T_1 + T_2) - m(O)$$

and

$$e_\psi(S, T_i) = \frac{g_i(X + S)}{g_i(X)}, \quad \text{for } i = 1, 2, 3.$$

Let $h \in \overline{\mathbb{F}}_q(E)$ be a function with divisor

$$\text{div}(h) = (T_1 + T_2) - (T_1) - (T_2) + (O). \quad (6.5)$$

Then, together with Equation (6.5), we get

$$\text{div}(f_3 / f_1 f_2) = m \text{div}(h),$$

so

$$f_3 = c f_1 f_2 h^m \quad \text{for some } c \in \overline{\mathbb{F}}_q^*.$$

We compose with $\psi$, and use the fact $g_i^m = f_i \circ \psi$ to obtain

$$g_3^m = (c f_1 f_2 h^m) \circ \psi = c (f_1 \circ \psi) (f_2 \circ \psi) (h^m \circ \psi) = c g_1^m g_2^m (h^m \circ \psi).$$
Next we take $m$-th roots. This only affects the equation by a constant factor. To demonstrate this, we proceed by taking divisors,

$$m \text{div}(g_3) = \text{div}(cg_1^m g_2^m (h^m \circ \psi))$$

$$= \text{div}(c) + \text{div}(g_1^m) + \text{div}(g_2^m) + \text{div}(h^m \circ \psi)$$

$$= 0 + m \text{div}(g_1) + m \text{div}(g_2) + \psi^* \text{div}(h^m)$$

$$= m \text{div}(g_1) + m \text{div}(g_2) + \psi^* \text{div}(h)$$

$$= m (\text{div}(g_1) + \text{div}(g_2) + \text{div}(h \circ \psi)).$$

and therefore,

$$g_3 = c' g_1 g_2 (h \circ \psi).$$

We can compute now

$$e_{\psi}(S, T_1 + T_2) = \frac{g_3(X + S)}{g_3(X)}$$

$$= \frac{g_1(X + S) g_2(X + S) h(\psi(X) + \psi(S))}{g_1(X) g_2(X) h(\psi(X))}$$

$$= \frac{g_1(X + S)}{g_1(X)} \frac{g_2(X + S)}{g_2(X)}$$

$$= e_{\psi}(S, T_1) e_{\psi}(S, T_2).$$

**Compatibility.** Let $\psi : E \to E$ be as before, and consider another isogeny $\delta : E \to E$. Let also $S \in \ker(\psi \circ \delta)$, and $T \in \ker(\delta)$. We want to relate $e_{\psi \circ \delta}$ and $e_{\psi}$.

For the definition of $e_{\psi}$, we take $f_T$ and $g_T$ as above. We consider now the definition of $e_{\psi \circ \delta}$. Let

$$\text{div}(F_T) = m'(T) = m'(O),$$

and

$$\text{div}(G_T) = (\psi \circ \delta)^* ((T) - (O))$$

$$= \delta^* \circ \psi^* ((T) - (O))$$

$$= \delta^* \text{div}(g_T)$$

$$= \text{div}(g_T \circ \delta).$$

We compute now,

$$e_{\psi \circ \delta}(S, T) = \frac{G_T(X + S)}{G_T(X)}$$
Theorem 6.3.1 Let $p$ be a prime, and $q = p^k$. Let $E/F_q$ be an elliptic curve, $\psi : E \to E$ be an endomorphism of $E$ of degree $m$ prime to $p$, and $\hat{\psi}$ its dual. Then there exists a pairing

$$e_\phi : \ker(\psi) \times \ker(\hat{\psi}) \to \mu_m,$$

with the following properties

1. **Bilinear:** $e_\phi(S_1 + S_2, T) = e_\phi(S_1, T)e_\phi(S_2, T)$, $e_\phi(S, T_1 + T_2) = e_\phi(S, T_1)e_\phi(S, T_2)$.

2. **Non-degenerate:** If $e_\phi(S, T) = 1$ for all $T \in \ker(\psi)$, then $T = O$.

3. **Compatible:** For another endomorphism $\delta : E \to E$, and $S \in \ker(\psi \circ \delta)$, $T \in \ker(\hat{\psi})$, we have

$$e_{\psi \circ \delta}(S, T) = e_\phi(\delta(S), T).$$

Remark. The generalized Weil pairing, as defined above, is a family of pairings, one for each endomorphism $\phi$, with $(\deg(\psi), p) = 1$. It is not hard to see that in the special case $\psi = [N]$, the resulting pairing is the Weil pairing. This justifies the name “generalized Weil pairing”.

### 6.4 Construction of new isomorphisms

In this section, we use the generalized Weil pairing to construct an isomorphism between a subgroup of $E(F_q)$ and a suitable group of roots of unity. Most of the ingredients are present. In particular, $e_\phi$ maps pairs of points to roots of unity, which form a group. We need to specialize the isogeny $\psi$, so that $\ker(\psi)$ is related to the group $E(F_q)$. The choice $\psi = 1 - \phi$ is natural, given of Theorem 5.8.6 and Equation (5.7).

Let $\hat{\psi} = 1 - \phi$, where $\phi$ is the $q$-th power Frobenius automorphism. Then we have $\hat{\psi} = 1 - \phi$, so from Equation (5.7) we have $\ker(\psi) = E(F_q)$. Also

$$\# \ker(\hat{\psi}) | \deg(\hat{\psi}) = \deg(1 - \phi) = \# E(F_q) = N,$$

where the divisibility comes from the fact that

$$\ker(\hat{\psi}) = \deg_s(\hat{\psi}),$$
and

\[ \text{deg}(\tilde{\psi}) = \text{deg}_s(\tilde{\psi}) \cdot \text{deg}_t(\tilde{\psi}). \]

Assuming that \( p \) does not divide \( N \), we have a bilinear, non-degenerate pairing

\[ e_\psi : E(\mathbb{F}_q) \times \ker(\tilde{\psi}) \to \mu_N. \]

We stress that this pairing exists and is bilinear and non-degenerate for any elliptic curve \( E \) and any finite field \( \mathbb{F}_q \).

6.4.1 Curves with \( a_q = 2 \)

In this section, we consider elliptic curves with trace of Frobenius \( a_q = 2 \). Let \( \phi \) be the \( q \)-th power Frobenius map, and \( P \in \ker(1 - \phi) = E(\mathbb{F}_q) \). We wish to find the point \( \tilde{\phi}(P) \). For that we consider the following

\[ (1 - \phi) \circ (1 - \hat{\phi}) = 1 - \phi - \hat{\phi} + [q]. \]

From the above observation, we have that

\[ (1 - \phi) \circ (1 - \hat{\phi})P = O. \]

Therefore,

\[ P - \phi(P) - \hat{\phi}(P) + [q]P = O, \]

which implies

\[ \tilde{\phi}(P) = [q]P. \]

We know that \( q + 1 - a_q = \#E(\mathbb{F}_q) \), therefore \([q]P = [a_q - 1]P\). Thus,

\[ \tilde{\phi}(P) = [a_q - 1]P. \]

Suppose now that the curve has \( a_q = 2 \). Then, for every point \( P \in E(\mathbb{F}_q) \) we have

\[ (1 - \tilde{\phi})P = O, \text{ i.e.,} \]

\[ E(\mathbb{F}_q) \subseteq \ker(1 - \tilde{\phi}). \]

Furthermore,

\[ \# \ker(1 - \tilde{\phi}) = \text{deg}_s(1 - \tilde{\phi}) \leq \text{deg}(1 - \tilde{\phi}) = \text{deg}(1 - \phi) = \#E(\mathbb{F}_q). \]

This implies that

\[ \ker(1 - \tilde{\phi}) = E(\mathbb{F}_q). \]

To summarize, for a curve \( E \) with trace of Frobenius \( a_q = 2 \), we have a pairing

\[ e_\psi : E(\mathbb{F}_q) \times E(\mathbb{F}_q) \to \mu_N, \]

where \( N = \#E(\mathbb{F}_q) \). Note that \( p \) (the characteristic) does not divide \( N = q - 1 \). Therefore from Theorem 6.3.1 it must be bilinear and non-degenerate.
6.4.2 A structure theorem

For the purposes of the section, we need to fix some more notation. The group \( E(\mathbb{F}_q) \) is isomorphic to \( \mathbb{Z}/n_1\mathbb{Z} \oplus \mathbb{Z}/n_2\mathbb{Z} \), with \( n_2 \mid n_1 \), and \( n_2 \mid q - 1 \). This means that \#\( E(\mathbb{F}_q) \) = \( N = n_1 n_2 \). We denote by \((T_1, T_2)\) a pair of generators of \( E(\mathbb{F}_q) \). Let \( P \) be a point in \( E(\mathbb{F}_q) \) of prime order \( r \). For the remainder of this chapter, we assume that \( n_1 = lr^k, r \not\mid l \) and that \( n_2 \not\mid r \), i.e., \( r \) does not divide \( n_2 \). This is typically the case in cryptography, as the point \( P \) is chosen to have very large order, close to the order of \( E(\mathbb{F}_q) \) itself. This means that in fact \( r \) is typically the exact power dividing \( N \). This is the only interesting case for any practical application. Theoretically, only minor things change if a higher power of \( r \) divides \( N \). We will indicate what happens in this case at the end of this section.

Since \( r \not\mid n_2 \), \((P)\) is contained in \((T_1)\). Our goal is to find a point \( P' \) such that \( e_\psi(P, P') \) is a primitive \( r \)-th root of unity.

**Lemma 6.4.1** There exist points \( T, S \in E(\mathbb{F}_q) \) such that \( e_\psi(T, S) \) is a primitive \( n_1 \)-st root of unity.

**Proof.** The image of \( e_\psi(T, S) \) as \( T \) and \( S \) range over \( E(\mathbb{F}_q) \) is a subgroup of \( \mu_N \), say equal to \( \mu_d \). Then it follows that for all \((T, S) \in E(\mathbb{F}_q) \times E(\mathbb{F}_q)\),

\[
1 = e_\psi(T, S)^d = e_\psi([d]T, S).
\]

The non-degeneracy of the \( e_\psi \) pairing implies that \([d]T = O\) for all \( T \in E(\mathbb{F}_q) \). In particular, if \( T = T_1 \) then it must be \( d = n_1 \). \( \square \)

**Lemma 6.4.2** The order of \( e_\psi(T_1, T_1) \) is divisible by \( r^k \).

**Proof.** Let \( T = [x_1]T_1 + [x_2]T_2 \) and \( S = [y_1]T_1 + [y_2]T_2 \) be one pair of points such that \( e_\psi(T, S) \) is a primitive \( n_1 \)-st root of unity, which exists by Lemma 6.4.1. Suppose now to the contrary, that \( r^k \) does not divide \( e_\psi(T_1, T_1) \). Then

\[
e_\psi(T, S) = e_\psi(T_1, T_1)^{x_1 y_1} e_\psi(T_1, T_2)^{x_1 y_2} e_\psi(T_2, T_1)^{x_2 y_1} e_\psi(T_2, T_2)^{x_2 y_2}
\]

Note now that the order of \( e_\psi(T_1, T_1) \) divides \( n_1 = lr^k \), but by assumption \( r^k \) does not divide it. Therefore, the order of \( e_\psi(T_1, T_1) \) divides \( lr^{k-1} \). Obviously, the orders of \( e_\psi(T_1, T_2), e_\psi(T_2, T_1), \) and \( e_\psi(T_2, T_2) \) divide \( l \). Thus we have,

\[
e_\psi(T_1, T_1)^{lr^{k-1}} = e_\psi(T_1, T_2)^{lr^{k-1}} = e_\psi(T_2, T_1)^{lr^{k-1}} = e_\psi(T_2, T_2)^{lr^{k-1}} = 1.
\]
Therefore,
\[ e_{\psi}(T, S)^{l^r-1} = 1, \]
which is a contradiction, since \( l^r-1 < n_1 \).

**Theorem 6.4.3** Let \( P' \in E(\mathbb{F}_q) \), be a point of order \( r^d \). Then, \( e_{\psi}(P, P') \) is a primitive \( r \)-th root of unity if and only if \( k < d + 1 \).

**Proof.** It is clear that \( e_{\psi}(P, P') \) is either a primitive \( r \)-th root of unity or 1. This is because
\[ e_{\psi}(P, P')^r = e_{\psi}([r]P, P') = e_{\psi}(O, P') = 1. \]
We recall that \( \langle P \rangle \), and \( \langle P' \rangle \) are subgroups of \( \langle T_1 \rangle \). It follows that \( P = [l^r-1]T_1 \) and \( P' = [l^r-d]T_1 \). Then we have
\[
e_{\psi}(P, P') = e_{\psi}(l^r-1]T_1, [l^r-d]T_1)
  = e_{\psi}(T_1, T_1)^{l^r-1}.
\]
Then Lemma 6.4.2 implies,
\[
e_{\psi}(P, P') = 1 \iff 2k - d - 1 \geq k \iff k \geq d + 1.
\]

The interesting case in cryptography is when the point \( P \) has very large order \( r \) (practically on the same order as \( q \)). For that reason, we state the following corollary.

**Corollary 6.4.4** Let \( P \in E(\mathbb{F}_q) \) be a point of order \( r \), such that \( r \neq \#E(\mathbb{F}_q) \). Then \( e_{\psi}(P, P) \) is a primitive \( r \)-th root of unity.

We want to emphasize that Theorem 6.4.3 is in sharp contrast with the properties of the Weil pairing. For the Weil pairing, \( w(P, P) = 1 \) for every \( P \in E[N] \). In our case, when \( k = 1 \) — which is the typical case in cryptography — the value \( e_{\psi}(P, P) \) is not trivial, and in fact is a primitive \( r \)-th root of unity. This eliminates a major obstacle of the Weil pairing approach: the point that makes \( e_{\psi}(P, \cdot) \) a primitive \( r \)-th root of unity is defined over \( \mathbb{F}_q \) (in the case of the Weil pairing it exists in a very large extension, unless the curve is supersingular). Furthermore, in our case, this point is known in advance.

We have the following theorem.
Chapter 6. The elliptic curve discrete logarithm problem

**Theorem 6.4.5** Let $P$ be a point in $E(\mathbb{F}_q)$ of prime order $r$, such that $r^2$ does not divide $N$. Then the map

$$V: \langle P \rangle \to \mu_r$$

given by

$$V(Q) = e_\psi(Q, P)$$

is a group isomorphism.

We now consider the case when $r^k$ is the exact power of $r$ that divides $N$. It follows from Theorem 6.4.3 that $e_\psi(P, P')$ is a primitive $r$-th root of unity, provided that $P'$ has order at least $r^k$ (and therefore cannot be equal to $P$, unless $k = 1$). So the task now is to find such a point. Here is a method to do that. First we compute a random point $P'' = (x, y) \in E(\mathbb{F}_q)$. Then we set

$$P' = \frac{N}{r^k} P''.$$

We claim that $P'$ is a point of order $r^k$ with high probability. Indeed, it is clear that $P'$ has order $r^d$ for some $d \leq k$. The probability that $P'$ has order exactly $r^k$ is equal to the probability that $r^k$ divides the order of $P''$. We can construct all the points of order divisible by $r^k$. Let $(T_1, T_2)$ be a pair of generators of $E(\mathbb{F}_q)$, so that

$$E(\mathbb{F}_q) = \frac{\mathbb{Z}}{l r^k \mathbb{Z}} T_1 \oplus \frac{\mathbb{Z}}{n_2 \mathbb{Z}}.$$

Then the points of order divisible by $r^k$ are

$$[i] T_1 + [j] T_2, \quad 1 \leq i < l r^k, \quad (i, r^k) = 1, \quad 1 \leq j \leq n_2.$$

Now, the number of integers from 1 to $l r^k$ that are coprime to $r^k$ are

$$l \varphi(r^k) = l r^{k-1} (r - 1).$$

Thus, the number of points in $E(\mathbb{F}_q)$ of order divisible by $r^k$ is

$$l r^{k-1} (r - 1) n_2.$$

We can compute the desired probability, now, as

$$\frac{l r^{k-1} (r - 1) n_2}{l r^k n_2} = 1 - \frac{1}{r}.$$
Theorem 6.4.6 Let $P$ be a point in $E(\mathbb{F}_q)$ of prime order $r$, such that $r^k$ is the exact power of $r$ dividing $N$. Then we can efficiently find a point $P' \in E(\mathbb{F}_q)$ of order $r^k$, and the map

$$V : \langle P \rangle \to \mu_r$$

given by

$$V(Q) = e_\psi(Q, P')$$

is a group isomorphism.

6.4.3 Computing the pairing

First remark

Lemma 6.4.7 The divisor $\text{div}(g_P)$ is defined over $\mathbb{F}_q$.

Proof. We wish to show that $(\text{div}(g_P))^\sigma = \text{div}(g_P)$ for every $\sigma \in \text{Gal}(\overline{\mathbb{F}_q}/\mathbb{F}_q)$. We recall that

$$\text{div}(g_P) = \sum_{\psi S = P} (S) - \sum_{\psi R = 0} (R).$$

By definition,

$$(\text{div}(g_P))^\sigma = \sum_{\psi S = P} (S^\sigma) - \sum_{\psi R = 0} (R^\sigma) = \sum_{\psi S = P} (S^\sigma) - \sum_{\psi R = 0} (R),$$

since $R^\sigma = R$. Consider any point $P'$ in the sum of $\text{div}(g_P)$. Then $\psi P' = P$. Also let $Q' = P'^\sigma$. Then we have,

$$\psi Q' = \psi (P'^\sigma) = (\psi P')^\sigma = P^\sigma = P,$$

which means that the point $Q'$ appears in the sum of $(\text{div}(g_P))^\sigma$. Also $\sigma$ is injective. In other words, $\sigma$ simply permutes the points in the sum, and thus, $(\text{div}(g_P))^\sigma = \text{div}(g_P)$. $\square$

We can use this lemma to show that the function $g_P$ has coefficients in $\mathbb{F}_q$.

Theorem 6.4.8 The function $g_P$ defined in Equation (6.3) is defined over $\mathbb{F}_q$.

Proof. It is in every case true that

$$\text{div}(g_P^\sigma) = (\text{div}(g_P))^\sigma.$$
CHAPTER 6. THE ELLIPTIC CURVE DISCRETE LOGARITHM PROBLEM

This fact combined with Lemma 6.4.7 implies that

\[
\text{div}(g_P^2) = \text{div}(g_P),
\]

that is, \(g_P\) is defined over \(\mathbb{F}_q\).

\[\square\]

Second remark

We recall, that the pairing is given by

\[
e_{\psi}(Q, P) = \frac{g_P(X + Q)}{g_P(X)},
\]

for any \(X \in E(\mathbb{F}_q)\). Suppose there is a function \(f \in \mathbb{F}_q(E)\), such that

\[
\frac{g_P(X + Q)}{g_P(X)} = \frac{f(X + Q)}{f(X)},
\]

(6.6)

for all \(X \in E(\mathbb{F}_q)\). Then clearly \(f\) can be used to compute \(e_{\psi}\). For such a function, if it exists, we have

\[
\frac{g_P(X + Q)}{g_P(X)} = \frac{f(X + Q)}{f(X)} \iff
\]

\[
\frac{g_P(X + Q)}{f(X + Q)} = \frac{g_P(X)}{f(X)} \iff
\]

\[
\frac{g_P}{f}(X + Q) = \frac{g_P}{f}(X) \iff
\]

\[
\frac{g_P}{f} \circ \tau_Q(X) = \frac{g_P}{f}(X) \iff
\]

\[
\tau_Q^* \left( \frac{g_P}{f} \right)(X) = \frac{g_P}{f}(X),
\]

for all \(X \in E(\mathbb{F}_q)\). This implies that

\[
\tau_Q^* \left( \frac{g_P}{f} \right) = \frac{g_P}{f}.
\]

This means, that \(g_P/f\) is fixed by \(\tau_Q^*\). From Theorem 5.8.3 part 3, we know that for every \(Q \in \ker(\psi) = E(\mathbb{F}_q)\), \(\tau_Q^*\) fixes exactly the field \(\psi^* \mathbb{F}_q(E)\). In other words, \(g_P/f\) is fixed by \(\tau_Q^*\) for every \(Q \in E(\mathbb{F}_q)\) if and only if

\[
\frac{g_P}{f} = \psi^* h = h \circ \psi, \quad \text{for some} \ h \in \mathbb{F}_q(E).
\]
Therefore, for any function $h \in \overline{\mathbb{F}}_q(E)$, the function

$$f = \frac{g_P}{h \circ \psi}$$

satisfies the conditions, i.e.,

$$\frac{g_P}{f} \in \psi^* \overline{\mathbb{F}}_q(E),$$

and therefore it is fixed by $\tau_Q^*$. Thus the function $f$ can be used in Equation (6.6) for the computation of $e_\psi$.

**The algorithm**

We give now an algorithm for computing the pairing $e_\psi$, in the case $\psi = 1 - \phi$. It is based on the following theorem, whose proof is given in the appendix.

**Theorem 6.4.9** Let $E/\mathbb{F}_q$ be an elliptic curve, with trace of Frobenius $a_q = 2$. Let $P \in E(\mathbb{F}_q)$ be a point of order $r$, and $Q \in E(\mathbb{F}_q)$, we have

$$e_\psi(Q, P) = \left( \frac{f_P(X)}{f_P(X + Q)} \right)^{(q-1)/r},$$

where $X \in E(\mathbb{F}_q)$ is any point such that $X \neq P$, and $F_P$ is a rational function with divisor

$$\text{div}(f_P) = r(P) - r(O).$$

Theorem 6.4.9 can be used to compute the value $e_\psi(Q, P)$. One first computes $f_P(X + Q)$ and $f_P(X)$ using repeated doubling. The point $X$ has to be chosen suitably, so that it does not appear in the support of the divisor of the functions that appear in the computation. The choice of $X$ can be done much in the same way as for the computation of the Weil pairing. Then the value $\left( \frac{f_P(X)}{f_P(X + Q)} \right)^{(q-1)/r}$ is computed using repeated squaring.

We have established an efficient reduction of the ECDLP for curves with $a_q = 2$ to the DLP in $\mathbb{F}_q^*$. The subexponential algorithm for the ECDLP in this case can be described as follows.

---

**Algorithm: Reduction (Curves with $a_q = 2$)**

**Input:** A point $P$ of order $r$, such that $r^2$ does not divide $\#E(\mathbb{F}_q)$, and a point $Q \in \langle P \rangle$

**Output:** An integer $\ell$ such that $Q = [\ell]P$

1. Compute $\alpha = e_\psi(P, P)$ and $\beta = e_\psi(Q, P)$.

2. Compute the discrete logarithm $\ell$ of $\beta$ to the base $\alpha$. 

---
3. Output $\ell$.

The correctness of the algorithm follows from the fact that the map

$$\langle P \rangle \mapsto \mu_r$$

$$Q \mapsto e_\psi(Q, P),$$

is a group isomorphism. This is given in Theorem 6.4.5. The efficiency of the reduction follows from Theorem 6.4.9. Therefore, step 1 is a polynomial time computation, and step 2 is a computation of a discrete logarithm in $\mathbb{F}_q$ which can be done in subexponential time using the index calculus method (see Chapter 4).

In the above algorithm, we clearly assumed that $e_\psi(P, P)$ is a primitive $r$-th root of unity. If this is not the case, and $r^k$ is the exact power of $r$ dividing $N$, then we can use Theorem 6.4.6, and find a point $P'$ such that $e_\psi(P, P')$ is a primitive root. The algorithm then has to be modified as follows.

**Algorithm: Reduction (Curves with $a_q = 2$)**

*Input:* A point $P$ of order $r$, such that $r^k$ is the exact power of $r$ that divides $\#E(\mathbb{F}_q)$, and a point $Q \in \langle P \rangle$

*Output:* An integer $\ell$ such that $Q = [\ell]P$

1. Compute a point $P' \in E(\mathbb{F}_q)$ of order $r^k$.
2. Compute $\alpha = e_\psi(P, P')$ and $\beta = e_\psi(Q, P')$.
3. Compute the discrete logarithm $\ell$ of $\beta$ to the base $\alpha$.
4. Output $\ell$.

### 6.5 Generalizations

The results as stated so far, give a subexponential algorithm for the ECDLP on elliptic curves with $a_q = 2$. They can, however, be generalized in two ways. First, if we require that the order $r$ of the point $P$ divides $q - 1$ (which is certainly true if $a_q = 2$), we still have a subexponential algorithm. Then we can generalize the attack to work in the case that $r$ divides $q^k - 1$ for some small $k$. This covers every possible reduction to the multiplicative group of an extension field (of $\mathbb{F}_q$).
6.5.1 The case \( r|q - 1 \)

In this section, we consider the first generalization, i.e., we assume that \( r|q - 1 \) which is equivalent to say that \( a_q \equiv 2 \pmod{r} \). We start from the pairing

\[
e_{\psi} : E(\mathbb{F}_q) \times \ker(1 - \hat{\phi}) \rightarrow \mu_N,
\]

which is bilinear and non-degenerate, as shown in Theorem 6.3.1. Since \( a_q \) is not necessarily equal to 2, we can not conclude that \( \ker(1 - \hat{\phi}) = E(\mathbb{F}_q) \) as in Section 6.4.1. We have, however, that \( \langle P \rangle \subseteq \ker(1 - \hat{\phi}) \). This is easily seen as follows. From the derivation of Section 6.4.1, and for any point \( Q \in \langle P \rangle \), we have

\[
\hat{\phi}(Q) = [a_q - 1]Q.
\]

Since \( a_q \equiv 2 \pmod{r} \), and \( Q \) is a point of order \( r \), we have

\[
\hat{\phi}(Q) = Q \quad \Rightarrow \quad Q \in \ker(1 - \hat{\phi}).
\]

For simplicity, we will only consider the case that no higher power of \( r \) divides \( N = \#E(\mathbb{F}_q) \) — which is the only interesting case in cryptography. Then we claim that \( e_{\psi}(P, P) \) is again a primitive \( r \)-th root of unity.

**Lemma 6.5.1** There exist a point \( S \in \ker(1 - \hat{\phi}) \), such that \( e_{\psi}(P, S) \) is a primitive \( r \)-th root of unity.

**Proof.** It is clear that \( e_{\psi}(P, S) \) is an \( r \)-th root of unity. Furthermore, as the point \( S \) ranges over \( \ker(1 - \hat{\phi}) \), the values \( e_{\psi}(P, S) \) are in a subgroup of \( \mu_N \), say \( \mu_d \). It follows that for all \( S \in \ker(1 - \hat{\phi}) \), we have

\[
1 = e_{\psi}(P, S)^d = e_{\psi}([d]P, S).
\]

The non-degeneracy of \( e_{\psi} \) then implies that \( [d]P = O \), i.e., \( r \) divides \( d \). It follows that the order of \( e_{\psi}(P, S) \) is exactly \( r \) for some point \( S \). \( \Box \)

As we pointed out in Section 6.4, we have \( \tilde{N} = \#\ker(1 - \hat{\phi})|N \). We also showed that \( r|\#\ker(1 - \hat{\phi}) \). We adopt the following notation: \( N = lr \), and \( \tilde{N} = l \tilde{r} \), with \( l \)|\( \tilde{r} \). Also, \( \ker(1 - \hat{\phi}) \) is the product of at most two cyclic groups, one of which contains \( \langle P \rangle \). If \( (S_1, S_2) \) is a pair of generators for \( \ker(1 - \hat{\phi}) \), it follows that the order of \( e_{\psi}(P, S_1) \) divides \( r \). If the order was 1, then it would violate the non-degeneracy of \( e_{\psi} \) (the argument is virtually the same as in Lemma 6.4.2 followed by Theorem 6.4.3 for \( k = 1 \)). Then, since \( P \in \langle S_1 \rangle \), it will be \( P = [l']S_1 \). Therefore,

\[
1 = e_{\psi}(P, P)^d = e_{\psi}(P, S_1)^{l'd},
\]

which implies that \( d \) has to be \( r \) (since \( r^2 \not|\tilde{N} \)). Therefore, we have the theorem.
**Theorem 6.5.2** Let $E / \mathbb{F}_q$ be an elliptic curve, $P \in E(\mathbb{F}_q)$ a point of prime order $r$ such that $r^2 \nmid N$, and assume that $a_q \equiv 2 \pmod{r}$. Then $e_\psi(P, P)$ is a primitive $r$-th root of unity.

We also note that the proof given in the appendix goes through in this more general case word by word. Therefore, the algorithm of the previous section works in the case $a_q \equiv 2 \pmod{r}$ as it is.

**Remark.** This remark concerns the size of $\ker(1 - \phi)$. It is clear that it is equal to the separable degree of $1 - \phi$. Also, we know that

$$\deg_s(1 - \phi) \deg_s(1 - \widehat{\phi}) = \deg(1 - \widehat{\phi}) = lr.$$ 

Since $1 - \phi$ is an isogeny, its inseparable degree is a power of the characteristic $p$. That means, $p|N$. In the case that $E$ is defined over a prime field $\mathbb{F}_p$, this would mean that $pr|N$, which is impossible (since $N < 2p$). That is, for curves $E / \mathbb{F}_p$, the endomorphism $1 - \widehat{\phi}$ is always separable. More generally, if $E$ is defined over $\mathbb{F}_q$, with $q = p^k$, and we assume that $\deg_s(1 - \phi) = p^e$ (clearly, $e \leq k$), then we have

$$p^e | N \quad \text{and} \quad p^e | q,$$

therefore we have $p^e | 1 - a_q$. So we get the condition

$$a_q \equiv 1 \pmod{p^e}.$$

### 6.5.2 The case $r|q^k - 1$

In this section, we comment about the case $r|q^k - 1$. We use the same construction as before, substituting $\psi = 1 - \phi^k$, where $\phi$ is the $q$ power Frobenius map as before. Then we get the following pairings.

- If $a_{q^k} = 2$ we have a pairing

$$e_\psi : E(\mathbb{F}_q^*) \times E(\mathbb{F}_q^*) \longrightarrow \mu_{N_k},$$

where $N_k = \#E(\mathbb{F}_q^*)$.

- More generally, if $a_{q^k} \equiv 2 \pmod{r}$, we have the pairing

$$e_\psi : E(\mathbb{F}_q^*) \times \ker(1 - \widehat{\phi^k}) \longrightarrow \mu_{N_k},$$

and as before, $\langle P \rangle \subseteq \ker(1 - \widehat{\phi^k})$.

The only complication here is to find a point $P'$ such that $e_\psi(P, P')$ is a primitive $r$-th root of unity.
Appendix

The computation of the generalized Weil pairing involves Galois cohomology. In this thesis we only need to use some basic facts that can be found in [64, Section X.1]. Standard references on the subject are [61, 60]. Let $E/\mathbb{F}_q$ be an elliptic curve, and let $\psi : E \to E$ be an isogeny. We start from the following exact sequence (see [64, Section X.1] or [6, Section V.2])

$$0 \to \ker(\psi) \to E(\mathbb{F}_q) \xrightarrow{\psi} E(\mathbb{F}_q) \to 0. \quad (6.8)$$

Taking $\text{Gal}(\overline{\mathbb{F}_q}/\mathbb{F}_q)$ cohomology, we obtain the following long sequence

$$0 \to E(\mathbb{F}_q) \cap \ker(\psi) \to E(\mathbb{F}_q) \xrightarrow{\psi} E(\mathbb{F}_q) \xrightarrow{\delta} H^1(G, \ker(\psi)) \to H^1(G, E(\mathbb{F}_q)) \xrightarrow{\psi} H^1(G, E(\overline{\mathbb{F}_q})).$$

where $G = \text{Gal}(\overline{\mathbb{F}_q}/\mathbb{F}_q)$. We can extract now the short exact sequence, sometimes called the Kummer sequence for $E/\mathbb{F}_q$

$$0 \to \frac{E(\mathbb{F}_q)}{\psi E(\mathbb{F}_q)} \xrightarrow{\delta} H^1(G, \ker(\psi)) \to H^1(G, E(\mathbb{F}_q))[\psi] \to 0, \quad (6.9)$$

where $H^1(G, E(\overline{\mathbb{F}_q}))[\psi]$ denotes the subgroup of $H^1(G, E(\overline{\mathbb{F}_q}))$ that is sent to the zero cocycle class by $\psi$. The connecting homomorphism $\delta$ is defined as follows. Let $P \in E(\mathbb{F}_q)$, and let $Q \in E(\overline{\mathbb{F}_q})$ such that $\psi(Q) = P$. Then a 1-cocycle representing $\delta(P)$ is given by

$$G \to \ker(\psi),$$

$$\sigma \mapsto Q^\sigma - Q,$$

that is

$$\delta(P)(\sigma) = Q^\sigma - Q.$$

If we assume now that $\ker(\psi) \subseteq E(\mathbb{F}_q)$, then the action of $G$ on $\ker(\psi)$ becomes trivial, and therefore

$$H^1(G, \ker(\psi)) = \text{Hom}(G, \ker(\psi)).$$

Furthermore, Hilbert's Theorem 90 provides the isomorphism

$$\frac{\mathbb{F}_q^*}{(\mathbb{F}_r)^*} \cong H^1(G, \mu_r).$$

Now, if we assume that $\mathbb{F}_q$ contains all the $r$-th roots of unity, then $\mu_r \subseteq \mathbb{F}_q$, therefore $G$ acts trivially on $\mu_r$, so

$$H^1(G, \mu_r) = \text{Hom}(G, \mu_r).$$
and we have the isomorphism

$$\delta_K : \mathbb{F}_q^r / (\mathbb{F}_q^r) \rightarrow \text{Hom}(G, \mu_r)$$
$$b \cdot (\mathbb{F}_q^r) \mapsto (\sigma \mapsto \beta^\sigma / \beta),$$

where $b \in \mathbb{F}_q^r$, $\beta \in \overline{\mathbb{F}}_q^r$, and $\beta^r = b$. In other words, for some $b \in \mathbb{F}_q^r$, $\delta_K(b)$ is a homomorphism from $G$ to $\mu_r$, and

$$\delta_K(b)(\sigma) = \frac{\beta^\sigma}{\beta}. \quad (6.10)$$

Then it can be shown (see [64, Section X.1] or [6, Section V.2]), that there exists a pairing

$$B : \frac{E(\mathbb{F}_q)}{\psi E(\mathbb{F}_q)} \times \ker(\psi) \rightarrow \frac{\mathbb{F}_q^r}{(\mathbb{F}_q^r)^r},$$

such that

$$e_\psi(\delta(S), T) = \delta_K(B(S, T)).$$

We note, that $\delta(S)$ is not a point in $\ker(\psi)$, and $\delta_K(B(S, T))$ is not an $r$-th root of unity. The above relation is to be interpreted as:

For any $\sigma \in G$, $e_\psi(\delta(S)(\sigma), T) = \delta_K(B(S, T))(\sigma). \quad (6.11)$

The crucial thing is that the bilinear pairing $B$ can be computed efficiently. In fact, if $S \neq T$, then

$$B(S, T) \equiv f_T(S) \pmod{(\mathbb{F}_q^r)^r},$$

where $f_T$ is a function with divisor

$$\text{div}(f_T) = r(T) - r(O).$$

If $T = S$, and $[2]T \neq O$, then we can use bilinearity to obtain

$$B(T, T) = f_T(-T)^{-1}.$$ 

More generally, for any point $X \neq T$ we have

$$B(S, T) = B(S + X - X, T) = B(S + X, T) B(-X, T)$$
$$= B(S + X, T) B(X, T)^{-1} = \frac{f_T(S + X)}{f_T(X)}.$$ 

We specialize now the isogeny $\psi$ to our case; that is $\psi = 1 - \phi$, where $\phi$ is the $q$-power Frobenius map. Furthermore, we deal with elliptic curves with trace of Frobenius $a_q = 2,$
i.e., \( \#E(\mathbb{F}_q) = N = q - 1 \). Our problem is the following: given points \( P, Q \in E(\mathbb{F}_q) \), with \( \#(P) = r \) compute \( e_\psi(Q, P) \). From Equation (6.11) we have

\[
e_\psi(\delta(S)(\sigma), P) = \delta_K(B(S, P))(\sigma),
\]

(6.12)

where \( \delta(S)(\sigma) = R^\sigma - R \) for some point \( R \) such that \( \psi(R) = S \).

We claim that if we choose \( \sigma \) to be the \( q \)-power Frobenius automorphism in \( G = \text{Gal} (\overline{\mathbb{F}}_q/\mathbb{F}_q) \), and \( S = -Q \), then we have

\[
\phi(R) = R^\sigma \text{ for any } R \in E. \tag{6.13}
\]

Then

\[
\psi(R) = S \implies
R - \phi(R) = S \implies
\phi(R) - R = -S \implies
R^\sigma - R = -S \implies
\delta(S)(\sigma) = Q.
\]

Thus, Equation (6.12) has become

\[
e_\psi(Q, P) = \delta_K(B(-Q, P))(\sigma), \tag{6.14}
\]

where \( \sigma \) is now fixed (and equal to the Frobenius automorphism).

It remains to compute \( \delta_K(B(-Q, P))(\sigma) \). We recall from Equation (6.10) that

\[
\delta_K(B(-Q, P))(\sigma) = \frac{\beta^\sigma}{\beta},
\]

where

\[
\beta^\sigma = B(-Q, P) = B(Q, P)^{-1} \equiv f_P(Q)^{-1} \pmod{(\mathbb{F}_q^*)^r}.
\]

We have,

\[
\frac{\beta^\sigma}{\beta} = \frac{\beta^q}{\beta} = \beta^{q-1}.
\]

Therefore, \( \delta_K(B(-Q, P))(\sigma) \) can be computed as

\[
\delta_K(B(-Q, P))(\sigma) = \left( \frac{f_P(X)}{f_P(X + Q)} \right)^{(q-1)/r},
\]

for any point \( X \in E(\overline{\mathbb{F}}_q), \ X \neq P \). Putting everything together, we have

\[
e_\psi(Q, P) = \left( \frac{f_P(X)}{f_P(X + Q)} \right)^{(q-1)/r}. \tag{6.15}
\]
Equation (6.1.5) can now be used to compute the value $e_\psi(Q, P)$. One first computes $f_P(X + Q)$ and $f_P(X)$ using repeated doubling. The point $X$ has to be chosen suitably, so that it does not appear in the support of the divisor of the functions that appear during the computation. This is done as described in [19]. The choice of $X$ can be done much in the same way as for the computation of the Weil pairing. Then the value $\left( \frac{f_P(X)}{f_P(X + Q)} \right)^{(q-1)/r}$ is computed using repeated squaring.

We note that Equation (6.1.5) shows that the pairing $e_\psi$ is the (multiplicative) inverse of the Tate-Lichtenbaum pairing that was used by Frey and Rück. Thus, our construction leads to essentially the same pairing, and therefore yields an equivalent reduction.
Chapter 7

Conclusions

The focus of this work was the discrete logarithm problem in finite fields and in the group of points of elliptic curves over finite fields. We conclude with some interesting open problems related to the subject of the thesis.

In the study of the discrete logarithm problem (as well as of integer factorization), the notion of smoothness is central. In the cases of integers and of polynomials over finite fields, the problem has been extensively studied. Integral ideals in number fields, however, have attracted less attention. In this context, the “size” of an ideal is measured by its norm. Thus, we want to estimate the number \( \psi(x, y) \) of ideals of norm at most \( x \), that factor into prime ideals whose norm is at most \( y \). The best results are lower bounds for \( \psi(x, y) \) and are contained in [9]. First order asymptotic results should be provable. Smooth ideals are relevant to the analysis of the algorithms in [8, 29] for the DLP in ideal class groups.

The discrete logarithm problem on elliptic curves still provides many challenges. Our reduction gives a subexponential algorithm in the case that the order of the cyclic subgroup \( \langle P \rangle \), where we work, divides \( q - 1 \). A slight generalization would establish the same result in the case \( \#\langle P \rangle | q^k - 1 \). This is also a necessary condition for any reduction to the DLP in the multiplicative group of a finite field. We believe that the idea of an efficient reduction can be exploited more. We propose to substitute the multiplicative group of a finite field by some other group, where the DLP is known to admit subexponential time algorithms. A possible candidate is the ideal class group of an algebraic number field, which we denote by \( C \). We mention that subexponential algorithms for the discrete logarithm problem in \( C \) are known [8, 29]. Such a reduction would involve the construction of an efficiently computable isomorphism

\[
\langle P \rangle \xrightarrow{\sim} C,
\]

95
where $P \in E(\mathbb{F}_q)$, and $C$ is the ideal class group of some number field. A connection between $E/\mathbb{F}_q$ and number fields comes from the fact that for an ordinary curve (non-supersingular), the endomorphism ring of $E$, denoted by $\text{End}(E)$, is an order in a quadratic imaginary field. Once a one-to-one homomorphism $\varphi : E \to \text{End}(E)$ has been established, then group cohomology can be employed to obtain an isomorphism between $E(\mathbb{F}_q)$ — the group of points defined over $\mathbb{F}_q$, which is the group of interest — and $\text{End}_{\mathbb{F}_q}(E)$, the ring of endomorphisms of $E$ that are defined over $\mathbb{F}_q$.

An important class of elliptic curves that is used in practical cryptosystems consists of the so-called subfield or Koblitz curves. Let $\mathbb{F}_q$ be a finite field, and $E/\mathbb{F}_q$ an elliptic curve defined over $\mathbb{F}_q$. It makes perfect sense to consider the group $E(\mathbb{F}_{q^n})$ for some $n > 1$. In this case, we call $E$ a subfield curve, since it is defined over a subfield of $\mathbb{F}_{q^n}$. A typical example are the Koblitz curves given by the equations

$$y^2 + xy = x^3 + 1 \quad \text{and} \quad y^2 + xy = x^3 + x^2 + 1,$$

both defined over $\mathbb{F}_2$, and then the groups $E(\mathbb{F}_{2^n})$ are considered. For subfield curves, the $q$-power Frobenius map $\phi$ is an endomorphism. Furthermore, starting with a point $P \in E(\mathbb{F}_{q^n})$, all the points $\phi(P), \phi^2(P), \ldots, \phi^n(P)$ are distinct. Gallant, Lambert, and Vanstone [22] and Wiener and Zuccherato [71] were able to explore this fact, and to modify Pollard's $\rho$ method to perform a random walk on sets of points rather than single points. This reduces the search space by a factor of $n$, and the total running time by a factor of $\sqrt{n}$. It is an interesting open question whether subexponential time algorithms exist for subfield curves. We note that the results of this thesis do not apply to all subfield curves. The ideas may, however, be useful, as the construction involves the Frobenius endomorphism which seems to be relevant to the problem.

A problem related to the index-calculus method, and in particular to Coppersmith's algorithm, is the following. Given a prime field $\mathbb{F}_p$ and an integer $n > 1$, are there irreducible polynomials of the form

$$f(X) = X^n + g(X), \quad \text{with} \quad \deg(g) \leq c \log n,$$

for some constant $c > 0$? Experimental results in [23] suggest that such irreducibles always exist, and in fact the constant $c$ is quite small. No proof of this is known. Let $f(X) \in \mathbb{Z}[X]$ be a polynomial over the integers, and denote by $\overline{f}(X)$ the polynomial over $\mathbb{Z}/p\mathbb{Z} = \mathbb{F}_p$ that we obtain by reducing the coefficients modulo $p$. Clearly $\overline{f}(X)$ can not be irreducible over $\mathbb{F}_p$ unless $f(X)$ is irreducible over $\mathbb{Z}$. If this is the case, then the way $\overline{f}(X)$ factors over $\mathbb{F}_p$ is determined by the way the prime $p$ splits in the algebraic
extension $\mathbb{Q}(\theta)$, where $\theta$ is a root of $f(X)$. This connection may be useful in establishing the existence of certain types of irreducibles such as those mentioned above.
Bibliography


INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

Bell & Howell Information and Learning
300 North Zeeb Road, Ann Arbor, MI 48106-1346 USA
800-521-0600
Logical Foundations of Shape-Based Object Recognition

Michael John Grüninger

Department of Computer Science
University of Toronto
Toronto, Ontario, Canada

A Dissertation submitted in conformity with the requirements
for the degree of Doctor of Philosophy

© Copyright by Michael John Grüninger 2000
The author has granted a non-exclusive licence allowing the National Library of Canada to reproduce, loan, distribute or sell copies of this thesis in microform, paper or electronic formats.

The author retains ownership of the copyright in this thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without the author’s permission.

L’auteur a accordé une licence non exclusive permettant à la Bibliothèque nationale du Canada de reproduire, prêter, distribuer ou vendre des copies de cette thèse sous la forme de microfiche/film, de reproduction sur papier ou sur format électronique.

L’auteur conserve la propriété du droit d’auteur qui protège cette thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation.

0-612-53693-9
Abstract

Logical Foundations of Shape-based Object Recognition

Michael Grüninger
Doctor of Philosophy
Graduate Department of Department of Computer Science
University of Toronto
2000

It has become evident in computational vision that there is a need to formalize the intuitions of existing vision systems. Any such formalization must allow the designer to explicitly reason about the capabilities and limitations of different representations and techniques. This thesis presents a logical axiomatization of shape-based object recognition in 2D scenes with occlusion and errors in edge detection, in a domain known as CardWorld. In particular, we introduce a family of eight theories which axiomatize the image and scene domain, polygonal shapes, depiction of shapes, reasoning about surface interiors, the depiction of surfaces as planar regions, the obscuring of surfaces in a scene, and the relationship between occlusion and depiction. We then introduce shape and depiction assumptions which allow us to restrict our attention to finite models of the CardWorld theories, and consequently specify propositional theories which are logically equivalent to the CardWorld theories. In the final part of the thesis, we consider the computational complexity of finding models of these propositional theories. We show that in general, these tasks are NP-hard, but by examining restricted classes of images, we can identify tractable subclasses of the problems.
Acknowledgements

I would like to thank my supervisor Ray Reiter for his support and guidance; his suggestions and criticisms have been invaluable in this work.

I would like to thank my committee – Allan Jepson, Hector Levesque, Alan Mackworth, and Craig Bouthilier – for having the perseverance to read my thesis.

Thanks to Mark Fox for his patience and support during my time in the Enterprise Integration Laboratory.

Special thanks to the members of Cafe Logique – Dale, Sheila, Javier, Manfred, Kathy, and Vincent. All of you provided so much encouragement through all these years: I couldn’t have done it without you. I’d also like to thank Joe Repka for our lunchtime conversations that had a profound impact on the methodology that I used in the thesis.

I gratefully acknowledge the financial support of the Department of Computer Science.
# Contents

1 Introduction to CardWorld ............................................ 1

1.1 Overview ........................................................................ 1

1.1.1 2D Object Recognition ........................................... 2

1.1.2 The CardWorld Theories ........................................... 3

1.1.3 Finite Model Assumptions ......................................... 4

1.1.4 Computational Complexity ......................................... 5

1.2 Methodology .................................................................. 5

1.3 Notation ........................................................................ 9

2 Object Recognition and Model Construction ......................... 10

2.1 Object Recognition and Constraint Satisfaction ................... 10

2.1.1 Constraints in Blocks Worlds ..................................... 11

2.1.2 Geometric Constraints ............................................. 12

2.2 Formal Approaches to Vision ......................................... 13

2.2.1 What is an Interpretation of an Image? ......................... 14

2.2.2 The Role of Inference .............................................. 15

2.2.3 Logic of Depiction .................................................. 16

2.3 Perceptual Organization and Causal Structure ................... 17

2.3.1 Semantics of Simplicity ............................................ 19

2.3.2 Grouping ............................................................ 20

2.4 Complexity Results for Object Recognition ....................... 21

3 Images, Scenes, and Depiction .......................................... 23

3.1 Model Theory of Images and Scenes ................................. 23

3.1.1 Languages for Images, Scenes, and Depiction ................ 23

3.1.2 Structures for the Kernel Theory ................................. 26

3.2 The Axioms in the Kernel Theory .................................... 32

3.2.1 Image Axioms .................................................... 32
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.2.2</td>
<td>Scene Axioms</td>
<td>32</td>
</tr>
<tr>
<td>3.2.3</td>
<td>Depiction Axioms</td>
<td>34</td>
</tr>
<tr>
<td>3.3</td>
<td>Characterization Theorems for $T_{kernel}$</td>
<td>38</td>
</tr>
<tr>
<td>3.3.1</td>
<td>Satisfiability of $T_{kernel}$</td>
<td>38</td>
</tr>
<tr>
<td>3.3.2</td>
<td>Depicted Kernel Theorem</td>
<td>41</td>
</tr>
<tr>
<td>3.3.3</td>
<td>Depicting Kernel Theorem</td>
<td>49</td>
</tr>
<tr>
<td>3.3.4</td>
<td>Characterizing Models of $T_{kernel}$</td>
<td>51</td>
</tr>
<tr>
<td>3.4</td>
<td>Assumptions about Images</td>
<td>52</td>
</tr>
<tr>
<td>4</td>
<td>An Axiomatic Theory of Shape</td>
<td>54</td>
</tr>
<tr>
<td>4.1</td>
<td>Geometric Relations for Polygonal Surfaces</td>
<td>54</td>
</tr>
<tr>
<td>4.1.1</td>
<td>Relative Alignment</td>
<td>56</td>
</tr>
<tr>
<td>4.1.2</td>
<td>Positional Relations</td>
<td>59</td>
</tr>
<tr>
<td>4.1.3</td>
<td>Contiguous Edges</td>
<td>61</td>
</tr>
<tr>
<td>4.1.4</td>
<td>Structures for $C_{polygon}$</td>
<td>62</td>
</tr>
<tr>
<td>4.2</td>
<td>Figure and Ground: Intuitions</td>
<td>65</td>
</tr>
<tr>
<td>4.2.1</td>
<td>Language of $T_{fg}$</td>
<td>65</td>
</tr>
<tr>
<td>4.2.2</td>
<td>Motivating Examples: What is a Surface?</td>
<td>66</td>
</tr>
<tr>
<td>4.2.3</td>
<td>Motivating Examples: Figure-Ground Assignments</td>
<td>67</td>
</tr>
<tr>
<td>4.3</td>
<td>Axiomatization of Figure and Ground</td>
<td>78</td>
</tr>
<tr>
<td>4.3.1</td>
<td>Structures for Figure and Ground</td>
<td>78</td>
</tr>
<tr>
<td>4.3.2</td>
<td>Characterization Theorems</td>
<td>88</td>
</tr>
<tr>
<td>4.3.3</td>
<td>Presenting $T_{fg}$</td>
<td>89</td>
</tr>
<tr>
<td>4.3.4</td>
<td>Satisfiability of $T_{fg}$</td>
<td>93</td>
</tr>
<tr>
<td>4.3.5</td>
<td>Axiomatizability Theorem for $T_{fg}$</td>
<td>96</td>
</tr>
<tr>
<td>4.3.6</td>
<td>Independence of Axioms in $T_{fg}$</td>
<td>103</td>
</tr>
<tr>
<td>5</td>
<td>Depiction Axioms for Figure and Ground</td>
<td>106</td>
</tr>
<tr>
<td>5.1</td>
<td>Extending the Language for Images</td>
<td>107</td>
</tr>
<tr>
<td>5.2</td>
<td>Motivating Examples for Depiction</td>
<td>108</td>
</tr>
<tr>
<td>5.2.1</td>
<td>Depiction of Alignment</td>
<td>111</td>
</tr>
<tr>
<td>5.2.2</td>
<td>Depiction and the between Relation</td>
<td>113</td>
</tr>
<tr>
<td>5.3</td>
<td>Structures for Depicted Surfaces</td>
<td>120</td>
</tr>
<tr>
<td>5.3.1</td>
<td>Relative Alignment and Depicted Edges</td>
<td>120</td>
</tr>
<tr>
<td>5.3.2</td>
<td>Depicting Figure and Ground</td>
<td>122</td>
</tr>
<tr>
<td>5.3.3</td>
<td>Characterization Theorems</td>
<td>128</td>
</tr>
<tr>
<td>5.4</td>
<td>The Axioms of $T_{\Delta}$</td>
<td>129</td>
</tr>
</tbody>
</table>
CONTENTS

3.5 Satisfiability of $T_{\Delta}$ .................................................. 133
3.6 Characterizing the Models of $T_{\Delta}$ ................................. 137
3.7 Conclusions ...................................................................... 146

6 Additional Theories for Scenes and Images .......................... 148
6.1 Interior of a Surface ............................................................. 148
   6.1.1 Some Geometry of Euclidean Planes ............................. 149
   6.1.2 Defining the Interior of a Surface ................................. 150
   6.1.3 Axiomatization of $T_{\text{interior}}$ ................................. 160
   6.1.4 Satisfiability of $T_{\text{interior}}$ ...................................... 164
   6.1.5 Characterization Theorems for $T_{\text{interior}}$ ............... 167
6.2 Images as Planar Graphs ..................................................... 173
6.3 Depicting the Interior of a Surface ...................................... 177
   6.3.1 Intuitions for Regions as Depicted Surfaces .................. 177
   6.3.2 Structures for Regions and Depicted Surfaces ................ 182
   6.3.3 Axiomatization of Depiction for $T_{\text{interior}}$ .............. 190
   6.3.4 Characterization Theorems for $T_{\text{interior}}$ ............... 194

7 A Theory of Occlusion ......................................................... 198
7.1 Obscuring Surfaces ............................................................ 198
   7.1.1 Intuitions for Obscuring Surfaces ................................ 199
   7.1.2 Structures for Obscured Scenes .................................. 203
   7.1.3 Axiomatizing $T_{\text{obscurers}}$ ...................................... 208
   7.1.4 Satisfiability Theorem for Obscuring Surfaces ............... 215
   7.1.5 Existence of Occluding Edges ..................................... 217
   7.1.6 Axiomatizability Theorem for $T_{\text{obscurers}}$ ............... 219
7.2 Occlusion and Depiction ..................................................... 222
   7.2.1 Structures for Occlusion ............................................ 222
   7.2.2 Occlusion Axioms .................................................... 227
   7.2.3 Satisfiability of $T_{\text{occ}}$ .......................................... 228
   7.2.4 Cues for Occlusion ................................................... 229
   7.2.5 Axiomatizability Theorem for Occlusion ...................... 235

8 Reasoning about Occlusion with No Errors in Edge Detection ... 239
8.1 Errors in Edge Detection: Overview ................................... 239
   8.1.1 Assumptions about Errors in Edge Detection ................. 240
   8.1.2 Preconditions for Propositional Theories ...................... 243
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>8.2</td>
<td>Regions and Connected Contours</td>
</tr>
<tr>
<td>8.3</td>
<td>Assumptions about Depicting Regions</td>
</tr>
<tr>
<td>8.3.1</td>
<td>Uniquely Depicting Regions</td>
</tr>
<tr>
<td>8.3.2</td>
<td>The Depicted Trichotomy Assumption</td>
</tr>
<tr>
<td>8.3.3</td>
<td>The Hole Region Existence Assumption</td>
</tr>
<tr>
<td>8.3.4</td>
<td>Summary and Implications</td>
</tr>
<tr>
<td>8.4</td>
<td>Assumptions of Region Existence</td>
</tr>
<tr>
<td>8.4.1</td>
<td>Existence of Depicting Regions</td>
</tr>
<tr>
<td>8.4.2</td>
<td>Lines and Region Existence</td>
</tr>
<tr>
<td>8.4.3</td>
<td>The Multiple Background Region Existence Assumption</td>
</tr>
<tr>
<td>8.5</td>
<td>Lacunate Errors and Region Existence</td>
</tr>
<tr>
<td>8.5.1</td>
<td>Bounded Region Existence Assumption</td>
</tr>
<tr>
<td>8.5.2</td>
<td>Occluding Surface Depiction Assumption</td>
</tr>
<tr>
<td>8.5.3</td>
<td>Summary and Implications</td>
</tr>
<tr>
<td>8.6</td>
<td>Characterizing NLEA</td>
</tr>
<tr>
<td>8.7</td>
<td>Applications of NLEA</td>
</tr>
<tr>
<td>9</td>
<td>Computational Framework for CardWorld</td>
</tr>
<tr>
<td>9.1</td>
<td>Closure Assumptions</td>
</tr>
<tr>
<td>9.1.1</td>
<td>Rectangular Shape Closure</td>
</tr>
<tr>
<td>9.1.2</td>
<td>Assumptions for Convex Surfaces</td>
</tr>
<tr>
<td>9.2</td>
<td>Simplifying the CardWorld Axioms</td>
</tr>
<tr>
<td>9.3</td>
<td>A Clausal Formulation</td>
</tr>
<tr>
<td>9.3.1</td>
<td>Finite Model Assumptions</td>
</tr>
<tr>
<td>9.3.2</td>
<td>A Useful Depiction Assumption</td>
</tr>
<tr>
<td>9.3.3</td>
<td>Poses for Occluded Surfaces</td>
</tr>
<tr>
<td>9.3.4</td>
<td>The CardWorld Clauses</td>
</tr>
<tr>
<td>9.4</td>
<td>Satisfying the Pose Clauses</td>
</tr>
<tr>
<td>10</td>
<td>Grouping Assumptions</td>
</tr>
<tr>
<td>10.1</td>
<td>Grouping</td>
</tr>
<tr>
<td>10.2</td>
<td>Methodology</td>
</tr>
<tr>
<td>10.3</td>
<td>Ambiguous Images</td>
</tr>
<tr>
<td>10.3.1</td>
<td>The Role of Ambiguous Lines</td>
</tr>
<tr>
<td>10.3.2</td>
<td>Intractability of Ambiguous Images</td>
</tr>
<tr>
<td>10.3.3</td>
<td>Tractable Classes of Images</td>
</tr>
<tr>
<td>10.4</td>
<td>Accidental Alignments</td>
</tr>
</tbody>
</table>
CONTENTS

10.4.1 Preliminary Definitions ......................................................... 343
10.4.2 Classes of Images for Accidental Alignments ............................... 348
10.4.3 Computational Framework for N.A.A ........................................ 352
10.4.4 Tractable Subclasses for Accidental Alignments ............................ 361
10.4.5 A Computational Framework for A.N.A.A .................................... 364
10.4.6 Tractable Subclasses for Abductive Accidental Alignments .............. 370

10.5 Region-restricted Accidental Alignments ....................................... 373
10.5.1 Definitions for Region-based Restrictions .................................... 374
10.5.2 Classes of Images for Region-Restricted Alignments ....................... 377
10.5.3 Computational Framework for R.N.A.A ....................................... 379
10.5.4 Tractable Subclasses for R.N.A.A ........................................... 384
10.5.5 Computational Framework for R.A.N.A.A .................................... 386

10.6 Summary ...................................................................................... 389

11 Summary and Future Work .................................................................. 390
11.1 Summary ..................................................................................... 390
11.2 Future Work .................................................................................. 391

A Hilbert's Axiomatization of Geometry ............................................... 394
A.1 The Axioms .................................................................................. 395
A.1.1 Incidence Axioms ....................................................................... 395
A.1.2 Order Axioms ............................................................................ 395
A.1.3 Congruence Axioms .................................................................... 395
A.1.4 Parallelism Axiom ....................................................................... 395
A.1.5 Completeness Axiom ................................................................... 395
A.2 Additional Relations ....................................................................... 396

A Index of Assumptions ........................................................................... 397
Chapter 1

Introduction to CardWorld

It has become evident in computational vision that there is a need to formalize the intuitions of existing vision systems. Any such formalization would allow the designer to explicitly reason about the capabilities and limitations of different representations and techniques.

This approach to the formalization of image understanding poses two challenges. First, we need to provide the foundations for a theory of vision that can define the capabilities of vision systems and why certain techniques fail when extended to other domains. For example, many systems have invisible, buried assumptions about their domain, not explicitly documented in publications, which must be rendered explicit if we are to identify principles on which vision systems are based.

The second aspect of image understanding which must be addressed is the role played by knowledge. We need to specify what knowledge is used in the vision process, and how it is used. We therefore need a theory that describes how one's knowledge of the world influences the interpretation an image.

In addressing these two challenges, this thesis provides some of the basic building blocks for constructing a formal framework for image understanding. We take a logical approach to vision that provides a framework for expressing the reasoned use of knowledge. The implicit assumptions that systems make about their domain are rendered explicit in a logical axiomatization. We consider image interpretation to be equivalent to model construction. In particular, we represent images, scenes, and depiction constraints as theories in first-order logic; interpretation of a specific image is then equivalent to finding models of the theories which also satisfy the sentences representing the image.

1.1 Overview

This thesis consists of three parts. The first part (Chapters 3-7) introduces the family of theories collectively referred to as the CardWorld Theories. These theories provide an axiomatization of images and depiction for 2D polygonal surfaces in scenes with occlusion and noise (errors in edge detection). The second part (Chapters 8 and 9) explore various assumptions which restrict the models of the CardWorld
Theories to be finite. This allows us to construct logically equivalent propositional theories. In the third and final part of the thesis (Chapter 10-11), we consider the computational complexity of finding models of these propositional theories. We show that in general, this task is NP-hard, but by examining the reasons for this intractability, we can formulate various assumptions about depiction and scenes which allow us to identify tractable subclasses of the problem. This also provides a semantic assumption-based approach to tractable reasoning - rather than identify syntactic classes of theories to achieve tractability, we can reason about the assumptions which a particular theory entails.

1.1.1 2D Object Recognition

The focus of this work (which we will call CardWorld) is the recognition of 2D planar objects in cluttered scenes with occlusion (such as Figure 1.1(a)) and errors in edge detection (such as Figure 1.1(b)). The goal of this thesis will be a logical characterization of this task; with the axiomatization, we will be able to provide an analysis of the role that various assumptions play in finding interpretations of an image.

For example, consider the image in Figure 1.1(a). One possible interpretation is that all of the regions in the image depict different surfaces. Another possible interpretation is that $R_1$ is a hole in the surface depicted by the region $R_2$, and that $R_6$ is background and does not depict any surface. The axiomatic theories we propose in this thesis must provide a characterization of all possible interpretations of an image.

We will also be making some restrictions on the domain. In particular, we will only be considering
2D surfaces which are opaque, and which are not self-occluding. We are not considering scene features such as colour, surface markings, or texture, and we are assuming that the projection of the scene into the image preserves metric information, such as the position of points.

1.1.2 The CardWorld Theories

In the first part of this thesis, we present eight theories to axiomatize intuitions about scenes, images, depiction, and occlusion for 2D polygonal surfaces. The relationships among these theories can be seen in Figure 1.2. We can consider each of $T_{\text{scene}}, T_{\text{kernel}}, T_f, T_{\Delta}, T_{\text{interior}}, T_{\Delta_{\text{interior}}, T_{\text{obsures}}, \text{ and } T_{\text{occ}}$ to be separate sets of axioms, or modules, which can be combined to form larger theories. The arrows in Figure 1.2 represents a dependency among the modules, in which one module requires the axioms in some other module. In particular, if $T_1$ and $T_2$ are two modules, and $T_1 \to T_2$, then any theory containing the set of axioms in $T_2$ must also contain the set of axioms in $T_1$. The complete set of axioms for a theory which includes some module is the union of the set of modules required by the module. In the theories we are considering, each module is a conservative extension of each of the theories which it requires.

For example, in Chapter 4, we will consider the axioms in $T_{\text{scene}} \cup T_{\text{kernel}} \cup T_f$; this is a conservative extension of the theory $T_{\text{scene}} \cup T_{\text{kernel}}$, which is considered in Chapter 3.
• $T_{\text{scene}}$: this theory axiomatizes the basic ontology of scenes within CardWorld, consisting of surfaces, edges, and points. (Chapter 3)

• $T_{\text{scene}} \cup T_{\text{kernel}}$: this theory extends $T_{\text{scene}}$ with the basic ontology of images (consisting of regions, lines, and pixels), as well as the basic intuitions about the depiction relation between scene objects and image objects. (Chapter 3)

• $T_{\text{fg}} \cup T_{\text{scene}}$: this theory extends $T_{\text{scene}}$ by axiomatizing intuitions about figure and ground for 2D polygonal surfaces, and provides a basic theory of shape. (Chapter 4)

• $T_{\Delta} \cup T_{\text{fg}} \cup T_{\text{kernel}} \cup T_{\text{scene}}$: this theory extends the previous theories by axiomatizing the depiction of the figure/ground relations in $T_{\text{fg}}$. (Chapter 5)

• $T_{\text{interior}} \cup T_{\text{fg}} \cup T_{\text{scene}}$: this theory extends the intuitions about figure and ground to define the notions of the interior of a surface (including any holes that the surface may have). (Chapter 6)

• $T_{\text{interior}}^{\Delta} \cup T_{\text{interior}} \cup T_{\Delta} \cup T_{\text{fg}} \cup T_{\text{kernel}} \cup T_{\text{scene}}$: this theory extends the previous one by axiomatizing the depiction of the interior of a surface and the relationship to properties of regions in an image. (Chapter 6)

• $T_{\text{obscures}} \cup T_{\text{interior}} \cup T_{\text{fg}} \cup T_{\text{scene}}$: this theory introduces the notion of obscuring surfaces within a scene. (Chapter 7)

• $T_{\text{occ}} \cup T_{\text{obscures}} \cup T_{\text{interior}}^{\Delta} \cup T_{\text{interior}} \cup T_{\Delta} \cup T_{\text{fg}} \cup T_{\text{kernel}} \cup T_{\text{scene}}$: this theory, which is the union of all of the modules in Figure 1.2, axiomatizes the relationship between the notion of obscuring surfaces and nondepiction of obscured scene objects. (Chapter 7)

The two most important modules within the CardWorld Theories are $T_{\Delta}$ and $T_{\text{occ}}$, which together formalize our intuitions about depiction and occlusion of 2D polygonal surfaces. We can consider the remaining modules to be those which are the necessary scaffolding to be able to provide these axiomatizations of depiction and occlusion.

All of the sets of axioms in Figure 1.2 except $T_{\text{fg}}$ are first-order theories; for $T_{\text{fg}}$, we will see that we need to introduce a second-order axiom to fully axiomatize the notion of connected sets of edges in a surface.

1.1.3 Finite Model Assumptions

Chapters 8 and 9 explore various assumptions which allow us to restrict our attention to finite models of the CardWorld Theories. The first set of assumptions deal with assumptions about errors in edge detection, the most important of which is the assumption that there are no undepicted scene objects which are not occluded: this eliminates those scene objects which are undepicted because of an error in
edge detection which leaves a "gap" in a line in the image. We show that this assumption is actually equivalent to the conjunction of ten other assumptions which can be made about the existence of regions, or whether a region depicts multiple or unique surfaces.

In Chapter 9 we make additional assumptions about the shapes of the surfaces which we will be considering. Together with the assumptions about no errors in edge detection, we show how we can restrict our attention to the finite models of the CardWorld theories. We also introduce the logically equivalent family of propositional theories whose models correspond to these finite models.

1.1.4 Computational Complexity

Having specified a family of propositional theories corresponding to the CardWorld Theories, we begin our investigation of the computational complexity of finding models of these propositional theories. Chapter 10 provides the basic theorem showing that the problem is in general NP-hard. However, we discover that the fundamental reason for the intractability arises from ambiguity in the depiction of edges and surfaces. We then introduce a set of depiction assumptions related to accidental alignment, which determine which depicted edges belong to the same surface. If the theory entails these assumptions, then the complexity of finding a model becomes polynomial.

1.2 Methodology

When presenting an axiomatization of some domain, we need some way of evaluating this axiomatization. In what sense is it correct? In what sense is it sufficient?

Figure 1.3 visualizes the methodology used in this thesis to evaluate the axiomatizations of the CardWorld theories.

The first aspect of this approach is to identify the primary intuitions in some domain. As we have already stated, in this thesis we are considering images and depiction for 2D polygonal surfaces in scenes with occlusion and noise (errors in edge detection). We thus have intuitions about concepts such as "surface", "region", "depiction", and "occlusion". These intuitions also restrict the scope of the axiomatic theories. In particular, we will only be considering 2D surfaces which are opaque, which are not self-occluding and which cannot inter-penetrate each other. We are not considering scene features such as colour, surface markings, or texture, and we are assuming that the projection of the scene into the image preserves metric information, such as the position of points.

For each CardWorld theory in this thesis, we will present an initial set of intuitions in the domain of interest. These will serve as informal requirements which get formally specified in the classes of structures, and later axiomatized in the theories themselves.

The second aspect of the methodology is the specification of some set of structures. These structures provides a rigorous mathematical characterization of the semantics of concepts in the domain. The
Figure 1.3: Methodology for the evaluation of axiomatic theories.
objective is to identify each concept with an element of some mathematical structure, such as a set or a set with additional structure; the underlying theory of the mathematical structure then becomes available as a basis for reasoning about the concepts and their relationships. Examples of structures include graphs, linear orderings, partial orderings, groups, fields, and vector spaces.

In particular, given the nonlogical lexicon in some language, structures are isomorphic to the extensions of the relations, functions, and constants denoted by the predicate symbols, function symbols, and constant symbols of the lexicon.

For each CardWorld theory in this thesis, we will define the class of structures. This will either be done by specifying some class of algebraic or combinatorial structures, or by extending classes of structures defined elsewhere in the thesis.

If we wish to model some domain, we want the necessary properties to be captured by the structures. Ideally, we want properties of the structures to be reflected by the properties of the corresponding concepts in the domain. These characteristics can be used to evaluate the adequacy of the intended structures. If some property is not captured, then we must make a decision about this property. If it is not necessary, then we can ignore it. If it is deemed necessary, then we must extend the characterization of the intended structures so that it includes some formalization of this property. For example, within CardWorld, we are only considering 2D surfaces which are opaque, which are not self-occluding, which cannot inter-penetrate each other. In addition, we are not considering scene features such as colour, surfaces markings, or texture. Since these properties are not considered within the scope of the work within this thesis, they are not formalized within any structures. Of course, if we wished to represent these properties, then we would need to extend the class of structures appropriately.

This relationship between the intuitions and the structures is, of course, informal, but we can consider the domain intuitions as providing a physical interpretation of the structures. In this sense, we can adopt an experimental or empirical approach to the evaluation of the class of intended structures in which we attempt to falsify these structures. If we can find some objects or behaviour within the domain which do not correspond to an intended structure, then we have provided a counterexample to the class of structures. In response, we can either redefine the scope of the class of structures (i.e. we do not include the behaviour within the characterization of the structures) or we can modify the definition of the class of structures so that they capture the new behaviour.

For example, physicists use various classes of differential equations to model different phenomena. However, they do not use ordinary linear differential equations to model heat diffusion, and they do not use second-order partial differential equations to model the kinematics of springs. If we wish to model some phenomena using a class of differential equations, we can use the equations to predict behaviour of the physical system; if the predictions are falsified by observations, then we have an incorrect set of equations. Similarly, in our case, we can use some class of structures to predict behaviour or characterize states of affairs; if there is no physical scenario in the domain which corresponds to these behaviours or
states of affairs, then we intuitively have an incorrect set of structures.

The final aspect of the methodology is the set of axiomatic theories, which are sets of sentences in some language using some logic. In this thesis, we primarily restrict ourselves to first-order languages, but as we shall see, there will be one case where we need to resort to sentences in second-order logic in order to fully axiomatize our intuitions.

Once we have specified the class of structures, we can formally evaluate an axiomatic theory with respect to this specification. In particular, we want to prove two fundamental properties:

- **Satisfiability**: every structure in the class is a model of the axiomatic theory.
- **Axiomatizability**: every model of the axiomatic theory is isomorphic to some structure in the class.

Strictly speaking, we only need to show that a model exists in order to demonstrate that a theory is satisfiable. However, in the axiomatization of domain theories, we need a complete characterization of the possible models. For example, since we are considering the domain of computer vision, to show that a theory is satisfiable, we need only specify an image and scene which together with the axioms are satisfied by some structure. The problem with this approach is that we run the risk of having demonstrated satisfiability only for some restricted class of images, scenes, or surfaces. For example, a theory of scenes and images may be shown to be consistent by constructing a satisfying interpretation, but the interpretation may require that there is no occlusion in the scene: although such a model may be adequate for such scenes, it would in no way be general enough for our purposes. We want to propose a comprehensive theory of 2D image interpretation, so we need to explicitly characterize the classes of images, scenes, surfaces, and other assumptions which are guaranteed to be satisfied by the specified structures. For example, in the chapter which proposes the theory of shape, the structures characterize the edges in a surface to be isomorphic to simple cycle graphs such that any simple cycle graph is isomorphic to the set of edges in a surface. Thus, the theory is satisfied not by an arbitrary set of edges, but rather by an entire class of graphs.

The purpose of the Axiomatizability Theorem is to demonstrate that there do not exist any unintended models of the theory, that is, any models which are not specified in the class of structures.

There are some classes of structures which are not definable within first-order logic, such as arithmetic and the class of connected graphs. In these cases there will be models of the first-order axiomatic theory which are not isomorphic to structures in the class. We then have two choices - accept these additional "unintended" models to be members of the class of structures, or use second-order sentences to eliminate the "unintended" models from the class of structures.

An example of the first choice is the theory of discrete linear orderings with initial and final elements. In addition to finite models, there are various countable models of this theory. However, since this theory axiomatizes the set of finite linear orderings, we could simply present the axioms and restrict our attention
to proving properties of the finite models if we wished.

We will see an example of the second choice in Chapter 4. The class of connected graphs is not first-order definable, even if we restrict ourselves to finite graphs. Since we will need to axiomatize surfaces as sets of connected edges, we will need to resort to a second-order axiom for connectedness.

*When presenting each of the CardWorld theories, we will prove the corresponding satisfiability and axiomatizability theorems.*

### 1.3 Notation

The work in this thesis relies heavily on concepts from model theory and graph theory. Although any specialized concepts from these fields will be introduced as needed, we first want to clarify some standard notation which will be used throughout:

- Structures are denoted by calligraphic font: $\mathcal{M}, \mathcal{N}, \mathcal{M'}, \mathcal{M}_1, \ldots$
- Classes of structures are denoted by $\mathcal{M}_i$.
- The domain of a structure $\mathcal{M}$ is denoted by $M$. Elements of a structure are denoted by *boldface* font, e.g. $e, I, s$. Within a structure $\mathcal{M}$, the denotation of a constant symbol $l_i$ in the language will be specified by $l_i$.
- The extension of relation $R$ in a structure $\mathcal{M}$ is denoted by 

$$\langle a_1, \ldots, a_n \rangle \in R$$

- Given a structure $\mathcal{M}$ and a variable assignment $\sigma$, then $\sigma(x)$ will represent the object in $\mathcal{M}$ denoted by the variable $x$.
- The distinction between constants and variables will be evident from the context. A letter which is not quantified in a sentence will refer to a variable *only* if it is part of an expression with an explicitly indicated variable assignment $\sigma$. All other appearances of a letter will indicate a constant within the language.
- In specifying a graph $G = (N, A)$, $N$ is the set of nodes in $G$ and $A$ is the set of arcs in $G$. The arc containing nodes $x$ and $y$ is denoted by $(x, y) \in A$.

In addition to notation, we will use the following conventions for diagrams:

- Surfaces (either isolated or together with other surfaces in a scene) will be shown as shaded figures.
- Diagrams which represent images will not be shaded.
Chapter 2

Object Recognition and Model Construction

There are three major themes within computer vision research which have motivated the work of this thesis:

- Approaches to image understanding which consider object recognition to be equivalent to the construction of a model of a first-order theory. The logic of depiction [Reiter and Mackworth 89] has been influential in this approach. This logical approach was in turn motivated by earlier work which considered object recognition to be a constraint satisfaction problem.

- Attempts at formalizing intuitions about perceptual organization and grouping ([Lowe 85], [Grimson 90], [Jacobs 88], [Feldman 97a]).

- Complexity analyses of object recognition tasks and algorithms ([Tsotsos 88], [Krousis and Papadimitriou 88], [Selman 89]).

In this chapter, we will consider related work in each of these three areas.

2.1 Object Recognition and Constraint Satisfaction

In this thesis, we will consider image interpretation to be equivalent to model construction. That is, we will represent images, scenes, and depiction constraints as theories in first-order logic; interpretation of a specific image will be equivalent to finding models of the theories which also satisfy the sentences representing the image. In this section, we consider related work in which interpretations of an image must satisfy some set of constraints. In the next section, we will consider approaches which extend constraint satisfaction to logical theories for object recognition.
2.1.1 Constraints in Blocks Worlds

A constraint-based approach to object recognition was first explored in the context of blocks worlds, in which the domain is simplified by explicit assumptions about the physical structure of the world.

Huffman and Clowes introduced the distinction between the scene domain and the image domain. The scene domain characterizes the physical structure of the (3D) world (e.g. occluding surfaces), while the image domain characterizes the depiction of the scene domain properties within a 2D image. Object recognition then becomes the problem of identifying the correspondence between scene features and image features.

Within blocks worlds, there are typically a restricted set of features:

<table>
<thead>
<tr>
<th>Scene</th>
<th>Image</th>
</tr>
</thead>
<tbody>
<tr>
<td>edge</td>
<td>line</td>
</tr>
<tr>
<td>vertex</td>
<td>junction</td>
</tr>
<tr>
<td>surface</td>
<td>region</td>
</tr>
</tbody>
</table>

A rigorous axiomatization of blocks world requires a description of how objects project to a line drawing, by studying how different kinds of points on the surface project, and building a catalog of the resulting junctions in the line drawing. In particular, Huffman and Clowes ([Huffman 71], [Clowes 71]) paid attention to the physical meaning of lines, and found useful constraints on types of lines at vertices. These constraints are used for interpretation of line drawing, so that recognition is equivalent to finding a consistent labelling for the lines in the image. Waltz extended this work in [Waltz 72] by incorporating an expanded set of scene features, and proposed a filtering algorithm that examined adjacent junctions in the image and discarded inconsistent labellings.

A crucial point is that possible configurations of labelled lines at junctions form a very small subset of all the combinations of assignments of labels to lines around junctions. Once all possible configurations at junctions are enumerated, the problem of interpreting line drawings can be reduced to a problem of assigning labels to lines consistently in the sense that the resultant configurations at junctions are all in the list of possible configurations. A combination of line labels for one junction type is referred to as a junction label. Each legal junction label implies which surfaces are connected at which edge in order to form that junction label. To assign labels to all lines in a picture means to choose one candidate for a spatial structure that the picture may represent. Thus the problem of interpretation of line drawings can be reduced to the problem of finding combinations of the labels that correspond to the spatial structures of the objects.

An additional problem which arises with this approach is physical realizability. Since the realizability of a line drawing is determined for a given interpretation, the realizability is undefined if the line drawing has no interpretation. However, a consistent labelling is only a necessary condition for a line drawing to be physically realizable; that is, a realizable drawing implies valid junction labelling, but a valid junction
labelling does not imply realizable drawing. Thus, some incorrect (unrealizable) line drawings can be labelled consistently.

Previous work has therefore treated interpretation as a two stage process ([Sugihara 86]): the first stage finds a consistent labelling, and the second stage filters out the labellings which are not physically realizable. The idea is that a consistent labelling merely gives a candidate for a spatial interpretation of the line drawing. Realizability is considered to be a property of interpretations, not drawings ([Kanade 80]). This reflects the loose use of the term “interpretation” and an insufficient notion of the constraints that are evident in the physical world. If we are working in the context of some underlying axiomatic theory of the scene and image, then an interpretation of a line drawing must necessarily be realizable, because it must satisfy the axioms of the theory. The problem of impossible figures that have consistent labellings is due to an incomplete axiomatization of these domains. The work which has been done in blocks world ([Mackworth 73], [Kanade 80], [Draper 81]) have all attempted to construct a more complete understanding of our intuitions of three dimensional structure. Some line drawings without consistent labellings may be realizable if the assumed constraints are relaxed (such as the trihedral assumption).

In order to remove nonsense labellings, we need to find a set of rules which labelled drawings must obey.

Having multiple interpretations is also an inevitable ambiguity so long as the theory is based solely on the basic constraints of objects concerning their realizability. In order to resolve the ambiguity, more knowledge and assumptions have to be incorporated.

Another aspect of 3D interpretations of line drawings is the problem of occlusion. In general, not all the objects in an image are fully visible; often one object will occlude another. The ability to reason with occlusion by predicting the hidden parts of these objects is an integral part of commonsense image understanding. However, not every junction in the image corresponds to a vertex in the scene. For instance, if A occludes B, then a line of A will intersect a line of B in the image, but this is only because of their position. In the interpretation of line drawings, these junctions are called T-junctions, and are given a separate “semantics” in the junction dictionary. It is usually assumed that the top of the T-junction corresponds to a nearer surface occluding another object, and that there is no constraint on the label of the stem of the T-junction.

2.1.2 Geometric Constraints

In more realistic domains, Grimson ([Grimson 90], [Grimson and Lozano-Perez 84], [Grimson and Lozano-Perez 87]) used the geometry of the shapes of objects to constrain the set of solutions for object recognition. In this approach, an interpretation of an image is a mapping from scene to image features which satisfies all of the geometric constraints.

Grimson identified various classes of geometric constraints. Unary constraints include constraints on the length of a feature – if the length of an image object is less than the length of the scene feature, then it is possible to assign the scene feature to the image feature.
Grimson also introduced several classes of binary geometric constraints. The first class of binary constraints is the angle constraint – if the angle between two linear image features is equal to the angle between two linear scene features, then it is possible to assign the scene feature to the image feature. We will see in Chapter 3 of this thesis how this class of constraints is axiomatized. The other two classes of binary constraints include the distance constraint between image and scene features, and the component constraint, which constrains the separation of two features using the range of components of a vector between them.

Interestingly, Grimson also makes claims concerning the completeness of these constraints. By complete, he means that in the absence of occlusion and error, the relative shape of two linear features can be reconstructed from the constraint measurements.

The other aspect of Grimson's approach is that these constraints ensure local consistency. To achieve global consistency, he introduces the interpretation tree, which is the set of possible assignments of scene features to image features. At each node of the tree, Grimson's algorithm applies the unary and binary constraints and prunes off inconsistent portions of the tree.

2.2 Formal Approaches to Vision

This section will review several approaches to formalizing vision. Several aspects of this formalization were presented in the introduction, but what exactly does this entail? The essential requirement for any formalism to be useful is that of correctness and completeness – it must find all and only the correct solutions. In order to prove that these properties hold for a given formalization of a task, we need to first find a specification of the problem. We then need a formalism that is adequate for the task to express these ideas.

In terms of Marr's levels ([Marr 82]), this thesis presents a computational theory of vision. Logic provides a formal framework for characterizing vision as a mapping from one kind of representation to another (namely from the image domain to the scene domain), and allows the abstract properties of this mapping to be defined precisely. Logic also allows us to uniquely define this mapping by representing the constraints which it must satisfy if we are to infer properties of the world from properties of images. The first goal is to therefore demonstrate the appropriateness and adequacy of a representation for a given task.

What constitutes an adequate representation of the image understanding process? Mackworth (in [Mackworth 87]) proposes several descriptive and procedural adequacy criteria, which define the soundness and completeness of the representation relation, that is, how adequately a given representation describes situations in the world. The adequacy of the representation refers not only to the expressiveness of the knowledge representation itself (descriptive adequacy), but also to the system's use of the knowledge (procedural adequacy). A representation must include all and only those interpretations
that map to the given image; many systems have failed to satisfy this correctness criterion in that they have excluded possible interpretations for an image (ambiguous interpretations), or included impossible interpretations (anomalous interpretations). It is only when the complete set of possible interpretations is completely specified that we can begin to then characterize which interpretations are plausible or preferred.

2.2.1 What is an Interpretation of an Image?

The semantics of a domain is the relationship between objects in the world and their images. We will be able to write programs which interpret pictures in some domain only if this relationship is specified. Brooks ([Brooks 81]) considered a system as “understanding” an image only if it could relate image features to a representation of the world that existed independently of image features. This relation is the depiction mapping, and its logical formulation is a necessary step in constructing a theory of image understanding.

In this approach, we need to determine how the properties of elements in the picture (image domain cues) can be related to the properties of elements in the scene (scene domain cues). Scene domain cues determine the physical meaning of objects in the scene, such as edge and surface configurations. Thus, edges are signified by labelled lines and surface orientation is represented by gradients in line drawings. Image domain cues refer to properties of features in the picture such as homogeneous regions, shapes of regions, intensity gradients, line segments, direction of lines, and parallelism of lines. An adequate representation scheme must maintain the distinction between these two domains. Many systems have failed to do this and have consequently confused the knowledge appropriate to these domains. The relation which defines how objects in the scene domain appear in the image is represented by the depiction relation or mapping. This includes specifying how various 3-D objects project onto the 2-D image plane, or what kinds of scene objects are depicted by image objects, as in [Reiter and Mackworth 89].

The first task is to formalize the scene, image, and mapping axioms. Again, it is essential to maintain the distinction between the picture and the scene. A picture (2D) consists of line segments which meet to form junctions, and which also form the regions into which the picture plane is divided. The scene (3D) consists of edges which meet each other at vertices and which separate surfaces: each surface may be thought of as a finite portion of some infinite plane.

It is tempting to say that the depiction relation defines the semantics of image understanding, since it defines which regions and lines in the image are meaningful in terms of our world knowledge about the scene. However, this view depends on what we consider an interpretation of an image.

One of the major contributions of the logic of depiction ([Reiter and Mackworth 89]) is the rigorous definition of an interpretation in the context of vision. As [Tsotsos 88] observes, little attention has been given to this problem. Current systems take an ad hoc approach, employing different representations
for interpretations as dictated by the specific problem domain. Only if the concept of an interpretation is precisely defined can we determine if an implementation is descriptively and procedurally adequate, finding all and only the interpretations allowed by world knowledge and the particular image.

The notion of interpretation is easier to specify in the context of line drawings, but there are still many different ways in which the term is used. The interpretation of a line drawing can be regarded as the assignment of possible labels, which represent the physical meaning of the image, to all lines in the drawing. Labels must be assigned so that all junctions may be included in the list of possible junction labels. The interpreted line drawing is a scene description because lines with labels represent features (edges of surfaces) of the scene.

Reiter and Mackworth define an interpretation of a particular image to be a logical model of the general knowledge (represented as scene and image domain axioms) and a description of how scene objects are depicted in the image (represented by the depiction mapping and a description of the image). This characterization of interpretations is the one which will be used throughout this thesis.

### 2.2.2 The Role of Inference

Work on recovering the physical properties of objects from images has led to the conclusion that the mechanisms for structuring the inferences in early vision must incorporate assumptions that hold abstractly over real world structure. Human perception must be exploiting the qualitative constraints provided by the structure of the scene in addition to quantitative models of the image information process. What is required, then, are theories of scene structure that capture the qualitative structure of our environment [Pentland 84], and describe how this structure is evidenced by regularities in the image.

Many researchers ([Pentland 84], [Hoffman and Richards 85]) have also noted that since sensory data is underconstraining, the additional knowledge required to derive assertions about the world means that perception is a process of inference. To have an adequate representation, we therefore need some set of axioms and rules of inference which generate descriptions of the legitimate primitive objects in the world and their possible attributes and their relationship, and a set of rules which say how these primitive objects combine into structured objects. To understand visual function — how one can infer knowledge about the world — it is necessary to have a model of the salient world structure and of how that structure evidences itself in the image. An ontology is a means of formally describing this set of structures; it is the agent's model of how the world is structured, how the agent "carves up" the world ([Pentland 84]). It is concerned with how our conceptual primitives (objects and relations) relate to each other and to the surrounding world. These conceptual primitives are not only objects. Regularities in the image also indicate structure in the environment, and recognition of these regularities allows the agent to infer consequences and anticipate events.

By rigorously specifying the ontology of a domain, we are able to understand images in terms of our world knowledge about the scene. This also allows us to determine what are legitimate assumptions to
make about the scene and image domains and what constraints they imply. By specifying the ontologies for different domains it also allows us to determine which structures in the scene project into which structures in the image. Witkin and Tenenbaum ([Witkin and Tenenbaum 83]) think of structural descriptions as providing an “alphabet soup” of descriptive chunks that have some direct semantic interpretation. Image understanding assimilates these chunks into more coherent interpretations.

The central goal of research into visual function is to first understand how the structure of the surrounding world is evidenced by regularities among the pixels of the image array, and then to understand how these regularities are mapped onto the predicates that constitute the primitive elements of cognition. However, the sensory data underdetermines the scene structure, because the pixels of the image array, by themselves, can determine nothing. Some knowledge of image formation and of how the world is structured is required in order to obtain any assertion about the viewed scene. This knowledge can be formally expressed in terms of the ontology axiomatization, which provides models of the world’s structure. A description that decomposes the image into constituents that capture regularity or coherence therefore provides descriptive chunks that act as “semantic precursors” in the sense that they deserve or demand explanation.

### 2.2.3 Logic of Depiction

Reiter and Mackworth have provided a precise, logical specification of the task addressed by the image understanding system Mapsee [Havens and Mackworth 83]. The domain of Mapsee is simplified hand drawn sketch maps of geographical regions (see Figure 2.1). There are two types of image objects: regions ($R_1, R_2$ in Figure 2.1) and chains ($A, B, C$ in Figure 2.1). There are four types of scene objects: land, water, river, and road. The task of Mapsee is to find interpretations of the image (i.e. correspondences between image and scene objects) using knowledge about real-world constraints.

The axiomatization in [Reiter and Mackworth 89] consists of three sets of first-order axioms: image axioms, scene axioms, and depiction axioms. The image axioms describe the map (i.e. the image) in terms of a set of image objects and the properties between them. Similarly, the scene axioms describe properties of the scene objects and constraints between them. The depiction axioms specify the mapping
between the image and scene domains, e.g. rivers in the scene are depicted by chains in the image.

The logic of depiction provides a framework for designing and implementing vision systems that are correct with respect to both task and algorithm levels. With this task level specification of the vision problem, we can analyze existing vision systems through a process of logical reconstruction. The basic premise of the work is that general knowledge of the image domain, the scene domain, and the depiction mapping can be expressed in first-order logic.

The logic of depiction proposed by Reiter and Mackworth has several limitations. In particular, Reiter and Mackworth make the Closure Assumption that all image domain predicates are known. This is a very strong assumption which is not valid in the case of ambiguity (since an image element may be any one of several elements). For a perfectly segmented line drawing, however, this assumption is valid. For complex scenes, image segmentation is highly ambiguous. The correspondence, or mapping, between regions of the image and physical objects in the scene is generally many to one. Boundaries between objects may not be distinguishable due to occlusion and the resolution of the image (McKeown et al 85). The assumption that regions in the segmented image directly correspond to objects in the scene is thus very ideal.

In addition, they make the unique names assumption, in which all image primitives are pairwise distinct, that is, if i and j are different constants denoting image primitives, then they denote different image elements. Due to this assumption, the logic of depiction also runs into trouble with occlusion, since in scenes with occlusion, there may no longer be a one-to-one correspondence between image objects and scene objects.

### 2.3 Perceptual Organization and Causal Structure

The fundamental characteristic of perceptual organization is the search for coherence and meaningful structure (Pragnanz) in the world. The basic premise is that the entities which result from the discovery of primitive spatio-temporal regularity correspond to causal events and processes in the world. To use this as a basis for doing primitive perceptual inference in a general and unified way, we must do better than enumerating a host of special case properties and relationships: we have to start with a general and unified expression derived from primitive properties of the world.

Lowe ([Lowe 85], [Lowe and Binford 87]) views segmentation as the division of the image into sets of related features. Perceptual organization is therefore defined by relations between image features, and the task of image segmentation is to distinguish the meaningful relationships between image elements from a background distribution of random alignments. In this account, low level vision focuses on discovering relations among elements replicating over space and time.

Previous methods for image segmentation have usually been derived from an idealized model of the world (such as the step edge model often used in edge detection). Lowe looks for meaningful groupings
in the image rather than for the image of some idealized feature in the world [(Lowe and Binford 87)]. In this framework, edge detection is then the detection of meaningful linear or curvilinear groups of points, where the values of the points have already been detected by earlier stages of the vision process.

Lowe and Binford define meaningfulness as the measure of how likely some grouping is to have arisen from an underlying physical relationship rather than through some accident of viewpoint or location. Groupings therefore establish relationships between elements of the image which plausibly correspond to existing structure in the scene. The interpretation of the image consists of mapping these primitive groupings to the physical entities. In this way, the image is segmented into meaningful pieces.

The depiction relation becomes the mapping from regularities in the image data to relevant structure in the environment based on the agent's ontology, or model of the world. The basic intuition underlying perceptual organization is that these regular relationships are so unlikely to arise by chance that they must reflect some underlying causal relationship.

All of the structural entities can be described in terms of spatiotemporal regularity or coherence: the relationships we notice are characterized by shapes, patterns, or configurations that replicate or continue with little or no change over an interval of time or space. All common manifestations of primitive structure can be characterized as a shape, pattern, or configuration that replicates or continues with little or no change over an interval of space or time. Good models are those in which relations persist as long as possible, either in time or according to the nature of the relationship. For example, if a set of points is collinear, then the desired models are those in which this collinearity persists as long as possible. Although many details need to be worked out with this approach, it seems to be able to express many of our commonsense intuitions about how the world is structured and how we perceive this structure.

The non-accidentalness argument says that what looks parallel or rigid really is parallel or rigid because the spurious appearance of parallelism is extremely unlikely to arise among causally unrelated curves. The meaning of structure follows directly from the non-accidentalness argument: when we strongly perceive a structural relationship (image), we are implicitly asserting that there is a corresponding causal relationship (domain).

Classes of meaningful alignments in an image also carry implications for the 3D structure of the scene. For example, given suitable assumptions on depiction, if two lines terminate at a certain point in the image, or three or more lines converge to a common point, then the edges that they depict will also terminate at a common point in 3-space; and curves which are parallel in the image are parallel in 3-space. In other words, a meaningful grouping in the image leads to inferences for the 3D structure.

We can view the imposition of this kind of structure as decomposing or explaining the image data in a particular way: when we perceive a smooth curve, we are replacing or augmenting our representation to the individual points on the curve with an explicit representation of a connected entity with a particular shape. When we notice parallelism or symmetry among a set of curves, we are replacing or augmenting our representation of the individual curves with a description of the common replicated shape and the
manner of its replication.

2.3.1 Semantics of Simplicity

The idea of finding the simplest description is found in several vision systems: for example, there are methods for image partitioning based on the simplest (most compact) description in some given descriptive formalism. However, it is difficult to define exactly what 'simple' means, and the success of this general approach is strongly dependent on the efficacy of the descriptive formalism that is assumed to be given. An adequate theory of vision must be able to specify the role which knowledge plays in image interpretation, and how to use particular quantitative methods, such as minimization, in a reasoned way which reflects the structure (semantics) which impose on our world.

Fischler and Bolles ([Fischler and Bolles 86]) argue that the underlying justification for Gestalt partitioning criteria is that such criteria provide the elements of a believable explanation of how the image was assembled from coherent parts. For example, partitioning can be viewed as an explanation of curve construction. To be believable, explanations must be concise, complete, and stable. The major problem with this proposal is that all of these characteristics are so high level that it is hard to find an operational definition, or a clear specification of what they mean.

Fischler and Bolles ([Fischler and Bolles 86]) also propose an alternative heuristic for forming explanations of images. A given set of data can be described by providing both the model parameters and the relationship of each data point to the model. A given body of data can be described in many possible ways by such a system, and choosing the simplest description corresponds to minimizing an objective function on the model components and the deviations of the data points from the instantiated model. When there are several different theories for representing some particular element of data, one theory is better than another if it yields a more compact explanation. Thus, what constitutes a good theory will always be dependent on our expectations about the world. The problem with this is that an observation which is preferred over another must have a more compact description in the representation scheme. Instead of preference being derived from the representation scheme (as described above), the representation scheme is dependent on the preference relation. It is obviously easier to change the preference relation among interpretations than to change the representation whenever a new observation needs explaining.

This approach appeals to Occam's razor: all else being equal, the most economical explanation should be chosen. Since the discovery of regularity reduces redundancy, the most regular description is the most economical, and therefore the best. But because the size of a description depends on the form in which it is represented, the argument is empty until the description terms are specified, and Occam's razor, by itself, offers no help specifying them.

Due to the coherence of matter, smoother or more regular shapes and structures are more likely to be observed than less smooth ones, and therefore given data that is consistent with a smooth surface, there is
a preference for the smooth surface. This argument has been made for the 3D interpretation of contours in [Barrow and Tenenbaum 81]. Brady and Yuille ([Brady and Yuille 83]) develop an extremum principle for determining 3D surface orientation from a 2D contour, which prefers symmetric surfaces. They base their measure on the following assumption: Contours that are the projections of curves in planes with large slant are most effective for deriving a 3D interpretation. Therefore, these interpretations are most are most preferred for shapes that are highly elongated in one direction. Given a contour, the extremum principle chooses an orientation in which the projected contour maximizes the compactness or symmetry of the surface. Thus, an ellipse is interpreted as a slanted circle and a parallelogram as a slanted square. Brady and Yuille also prove that this extremum principle necessarily interprets skewed symmetries as oriented real symmetries, which Kanade ([Kanade and Kender 83]) proposed as a heuristic assumption.

2.3.2 Grouping

The motivation behind grouping in image interpretation is that local geometric relations in the image can be used to identify structure in the scene. This is usually justified by some non-accidentalness principle – it is unlikely that two lines in the image satisfy some relation when the edges corresponding to these lines do not satisfy the corresponding relation in the scene. Such geometric relations include collinear and parallel lines, intersecting lines, and proximity. In particular, [Jacobs 88] considers the following condition:

Suppose one extracts convex sets of connected edges in an image of a cluttered scene. For each pair of such sets, can one derive an estimate of the likelihood that both sets arose from the same object?

Huttenlocher and Wayner ([Huttenlocher and Wayner 92]) extended this approach by a more sophisticated representation of convexity constraints. Both of these pieces of work have served as the basis for intuitions about shape representation in Chapter 4 of this thesis.

Previous approaches (e.g. [Grimson 90], [Lowe 85], [Jacobs 88]) have considered particular arrangements of lines that are “nonaccidental” insofar as they are unlikely to be depicted in the image unless they actually exist in the scene. In this probabilistic approach, nonaccidental features have a high likelihood ratio in favour of conjecturing that the image objects depict related scene objects.

An alternative approach ([Jepson and Richards 92], [Richards et al 96]) defines “regular” configurations, which are constructed from generic ones by removing degrees of freedom. Thus, preferred configurations have fewer degrees of freedom, and less preferred configurations have more degrees of freedom.

The work of Feldman ([Feldman 97a,97b,97c]) is most similar to the approach taken in this thesis. In particular, [Feldman 97a] proposes a logical approach to grouping. In this approach, he defines lattice structure over the set of models for an image and scene. Within this lattice, a model $M_1$ is more
preferred than a model $\mathcal{M}_2$ if $\mathcal{M}_1$ satisfies some "regularity" in the scene and $\mathcal{M}_2$ does not satisfy this regularity. In general, regularities are specified by the user, and would include standard intuitions such as collinear and parallel lines. Thus, models in which there are no accidental alignments (i.e. image elements can be grouped as depicting the same scene object) will be preferred over models in which there are accidental alignments (i.e. the image elements depict unrelated scene objects).

Feldman considers lattices of arbitrary models, rather than the models of a specific axiomatic theory, although he does apply these notions to a simple "Dots World" containing configurations of points in the plane.

Previous work ([Jacobs 88]) has also been aimed at designing local algorithms that find possible groupings. The groupings generated in this way are not guaranteed to be globally consistent: it is even unclear what it means for a grouping to be consistent. One of the tasks of an axiomatic theory of object recognition is to clarify these intuitions about the consistency of grouping.

### 2.4 Complexity Results for Object Recognition

Many of the complexity results for object recognition have considered the complexity of specific algorithms. For example, [Grimson 90] shows that the complexity of his interpretation tree algorithm for solving geometric constraints is exponential. However, in this thesis we will be more concerned with the complexity of the image interpretation task itself.

Tsotsos ([Tsotsos 88]) provided early work in establishing NP-completeness results for vision tasks. In particular, he showed that the task of bottom-up visual search is NP-complete. The general task of visual search seeks to find the subset of the image that matches some set of scene features. In bottom-up visual search, the set of scene objects which are depicted in the image are not known a priori.

Kirousis and Papadimitriou ([Kirousis and Papadimitriou 88]) provide complexity results for the blocks world domain. Given the line drawing, the problem is to decide whether it is the projection of the visible part of a set of opaque polyhedra. They show that this problem is NP-complete, even in the case of trihedral scenes (no four planes share a point) without shadows and cracks. They also introduce the class of orthohedral scenes, for which there is a linear algorithm to determine whether or not an image can be labelled. A scene is orthohedral if it has the following property: All planes in it are normal to one of the axes, so that each edge is parallel to one of the axes. The NP-completeness proof uses the reduction to Planar 3SAT [Lichtenstein 82], which is the same class of formulae which will be used in the complexity results in the final chapter of this thesis.

Selman ([Selman 89]) provides a complexity analysis for the Mapsee domain using Reiter and Mackworth's logic of depiction. He shows that although each sketch map has a trivial interpretation (all regions are interpreted as land and all chains depict roads), finding a nontrivial interpretation by completing a partial interpretation is NP-hard. As with [Kirousis and Papadimitriou 88], Selman also uses
the reduction to Planar 3SAT. Of particular interest, Selman shows that the intractability is due to the fact that there are rare interpretations that are difficult to compute. Interpretations in this class have what he calls river-road conflicts, which are structures consisting of two chains $x, y$ which intersect in a T-junction, and such that $x$ depicts a river and $y$ depicts a road. For interpretations which do not have these conflicts, there exists a linear algorithm to find a complete nontrivial interpretation.

Another related piece of work is [Cooper 92], who proposes parsing algorithms for interpreting images of cluttered scenes and partial occlusion. Cooper considers a grammar which generates possible images: in this approach, image interpretation parses the image using this grammar to find the corresponding scene. He also shows that if the scenes allow cycles of occluding objects, then the complexity of finding a parse of the image is NP-hard. However, Cooper does not provide an axiomatic theory for this task, and this makes it difficult to see how his results can be generalized to different classes of images and scenes.
Chapter 3

Images, Scenes, and Depiction

The focus of this work (which we will call CardWorld) is the recognition of 2D planar objects in cluttered scenes, based on the shapes of the objects in the scene. The goal will be a logical characterization of this task: with the axiomatization, we will be able to provide an analysis of the role that various assumptions play in finding interpretations of an image.

The first step in this approach will be the axiomatization of the ontology for image and scene objects and the depiction relation between these two kinds of objects. In this chapter, we present the two basic axiomatic theories which will be needed to build the entire set of CardWorld theories. We provide a definition of the class of structures which formalize our intuitions about scenes, images, and depiction, and then prove characterization theorems about these structures and the axiomatic theories.

3.1 Model Theory of Images and Scenes

We begin by recognizing that there are two fundamental classes of objects in the ontology of any vision problem - image and scene. We will first consider in depth the language and structures that we will be using for images and scenes. In the next section, we will consider assumptions over these domains, and axioms for scene objects and the depiction relation between the scene and image.

3.1.1 Languages for Images, Scenes, and Depiction

The language which we will adopt for images is a first-order language with equality. The nonlogical lexicon is:

\[ L_{image} = \{ R_i, L_j, Q_k, image\_object(x), region(r), line(l), pixel(q), in(x, y) \} \]

- There are three mutually exclusive sets of constants which intuitively will denote image objects – regions, lines, and pixels. \( R_i \) is a countably infinite set of constant symbols denoting regions in an

\[ ^1\text{Throughout this work, we are using the standard interpretation for the equality predicate } =. \]

23
image. $L_j$ is a countably infinite set of constant symbols denoting lines in an image, and $Q_k$ is a countably infinite set of constant symbols denoting pixels in an image.

- The unary predicate symbol \textit{region} denotes the relation which defines the set of regions in an image.
- The unary predicate symbol \textit{line} denotes the relation which defines the set of lines in an image.
- The unary predicate symbol \textit{pixel} denotes the relation which defines the set of pixels in an image.
- The binary predicate symbol \textit{in} which denotes the relation which defines a composition hierarchy over image objects.

Thus for lines and regions the intended interpretation is that line $l$ is contained in region $r$, and for pixels and lines the intended interpretation is that pixel $q$ is contained in line $l$.

Additional image relations will be defined in later theories within the thesis; the languages for these theories will all be expansions of $L_{\text{image}}$.

\textbf{Example:} Given the image in Figure 3.1, there is a structure $M$ in $L_{\text{image}}$ such that

$$\{r_1, r_2, l_1, l_2, l_3, l_4, l_5, l_6, l_7, l_8, q_0, q_1, q_2, q_3, q_4, q_5, q_6, q_7, q_8, q_9, q_{10}\} \subseteq M$$

and

$$\{r_1, r_2, l_1, l_2, l_3, l_4, l_5, l_6, l_7, l_8, q_0, q_1, q_2, q_3, q_4, q_5, q_6, q_7, q_8, q_9, q_{10}\} \in \text{image\_object}$$

$$\{r_1, r_2\} \in \text{region}$$

$$\{l_1, l_2, l_3, l_4, l_5, l_6, l_7\} \in \text{line}$$

$$\{q_0, q_1, q_2, q_3, q_4, q_5, q_6, q_7, q_8, q_9, q_{10}\} \in \text{pixel}$$

It should be noted that while this is the complete set of regions and lines in the image, it is not the complete set of pixels for the above image. In practice, this would depend on the resolution of the edge-detection algorithms that produce the image. The essential idea here is that for each line there exists a set of pixels which are its elements. We will see later that every line must contain at least two pixels as elements. \hfill \Box

As another example, in Figure 3.10(f), the regions are denoted by the constants $r_1, r_2, r_3$, and the lines are denoted by the constants $l_1, \ldots, l_{14}$.

The language which we will adopt for scenes is a first-order language with equality; the nonlogical lexicon is:

$$L_{\text{scene}} = \{S, E, P, \text{scene\_object}(x), \text{surface}(s), \text{edge}(e), \text{point}(p), \text{part}(x, y)\}$$
There are three mutually exclusive sets of constants which intuitively will denote scene objects – surfaces, edges, and points. $S_i$ is a countably infinite set of constant symbols denoting surfaces, $E_j$ is a countably infinite set of constant symbols denoting edges, and $P_k$ is a countably infinite set of constant symbols denoting points.

- The unary predicate symbol $\text{surface}$ denotes the relation which defines the set of surfaces.
- The unary predicate symbol $\text{edge}$ denotes the relation which defines the set of edges.
- The unary predicate symbol $\text{point}$ denotes the relation which defines the set of points.
- The binary predicate symbol $\text{part}$ denotes the relation which defines a composition hierarchy over scene objects.

Additional scene relations will be defined in later theories within the thesis: the languages for these theories will all be expansions of $L_{\text{scene}}$.

Example: Given the surfaces in Figure 3.2, there is a structure $\mathcal{M}$ in $L_{\text{scene}}$ such that

$$\{s_1, s_2, e_1, e_2, e_3, e_4, e_5, e_6, e_7, p_1, p_2, p_3, p_4, p_5, p_6, p_7\} \subseteq M$$

and

$$\{s_1, s_2, e_1, e_2, e_3, e_4, e_5, e_6, e_7, p_1, p_2, p_3, p_4, p_5, p_6, p_7\} \in \text{scene.object}$$

$$\{s_1, s_2\} \in \text{surface}$$

$$\{e_1, e_2, e_3, e_4, e_5, e_6, e_7\} \in \text{edge}$$

$$\{p_1, p_2, p_3, p_4, p_5, p_6, p_7\} \in \text{point}$$

It should be noted that while this is the complete set of surfaces and edges in the structure, it is not the complete set of points. In practice, this would depend on the granularity required by the tasks and chosen by the user. The essential idea here is that for each edge there exists a set of points which are its elements. However, there are some guidelines provided by the axioms of the CardWorld theories; we will see later in this chapter that every edge must contain at least two points as elements, and in Chapter 7 we will find the axiom stating that points are scene objects that cannot be partially occluded. $\square$
The language which we will adopt for depiction is a first-order language with equality. The nonlogical lexicon is:

\[ \mathcal{L}_{kernel} = \mathcal{L}_{image} \cup \mathcal{L}_{scene} \cup \{ \Delta(i,s) \} \]

- The binary predicate symbol \( \Delta \) denotes the relation which specifies the depiction relation between image objects and scene objects. Intuitively, \( \Delta(i,s) \) means that image object \( i \) depicts scene object \( s \).

Example: Given the image in Figure 3.1 and the surfaces in Figure 3.2, there is a structure \( \mathcal{M} \) in \( \mathcal{L}_{kernel} \) such that

\[ \{ (r_1,s_2), (r_2,s_1), (l_3,e_2), (l_3,e_7) \} \in \Delta \]

Notice the line \( l_3 \) depicts an edge from each surface: intuitively, this arises from abutting surfaces in a scene. \( \square \)

### 3.1.2 Structures for the Kernel Theory

The relation \( \text{in} \) over image elements and the relation \( \text{part} \) over scene elements provide additional structure for these sets of elements. For example, surfaces always contain edges and points, and edges in turn contain points (see Figure 3.2). Edges and points can never exist in isolation – they must always be part of a surface. There are also some additional constraints – edges and points cannot be part of multiple surfaces.

In the image domain, regions always contain lines and pixels, and lines contain pixels. However, it is possible for lines to exist which are not contained in a region, and pixels can exist which are not contained in a line. (See Figure 3.1)
Figure 3.3: Graph representations of the incidence structures for scene elements in Figure 3.2.

Figure 3.4: Graph representations of the incidence structures for image elements in Figure 3.1.

To formally capture these intuitions, we will first need to define a few classes of combinatorial structures which will be the building blocks.

**Definition 3.1** A tripartite incidence structure is a tuple $G = (\Omega_1, \Omega_2, \Omega_3, I)$, where $\Omega_1, \Omega_2, \Omega_3$ are pairwise disjoint sets with

$I \subseteq \Omega_1 \times \Omega_2 \cup \Omega_1 \times \Omega_3 \cup \Omega_2 \times \Omega_3$

We will say that elements in $\Omega_1$ have type 1, elements in $\Omega_2$ have type 2, and elements in $\Omega_3$ have type 3.

Two elements of $G$ that are related by $I$ are called incident.

**Definition 3.2** Let $G = (\Omega_1, \Omega_2, \Omega_3, I)$ be a tripartite incidence structure. A flag of $G$ is a set of elements of $\Omega_1 \cup \Omega_2 \cup \Omega_3$ that are mutually incident. A flag $F$ is maximal if there is no element $x \in \Omega \setminus F$ such that $F \cup \{x\}$ is also a flag.

The incidence structure $G$ has a rank $r$ iff all maximal flags has exactly $r$ elements.

Using these definitions, our intuitions tell us that scene elements should form incidence structures of rank 3, since all scene objects should be part of some surface, i.e., there should not exist any isolated edges or points. On the other hand, not every image object must be an element of a region, since there...
can exist isolated lines and pixels. Image elements should therefore form incidence structures with no rank, since not all maximal flags contain the same number of elements. In Figure 3.1, the region $r_1$ contains the line $l_1$ and the pixel $q_1$, and so a maximal flag associated with $r_1$ contains three elements. However, the maximal flag associated with the line $l_2$ contains only two elements, and the maximal flag associated with $q_3$ contains only one element.

We can represent the incidence structures as graphs, in which the image or scene elements are nodes of the graph, and there is a path between two nodes iff the elements are incident. The graph representation of the incidence structure for the image in Figure 3.1 is given in Figure 3.4, and the graph representation of the incidence structure for the scene in Figure 3.2 is given in Figure 3.3.

**Definition 3.3** Suppose that $M$ is a structure in $L_{\text{kernel}}$.
An image substructure $I(i)$ for an image object $i$ is the following set:

$$I(i) = \{ x : x \in M, (x, i) \in \text{in} \}$$

Intuitively, the image substructure for an image object is the set of image objects which are contained in it. In terms of the incidence structures, the image substructure for an element is the set of image elements in the flag of the element. In Figure 3.1, the image substructure for $r_1$ is $\{l_1, l_2, l_3, q_1, q_2, q_4\}$, the image substructure for $l_7$ is $\{q_0, q_8\}$, and the image substructure for $q_3$ is empty.

**Definition 3.4** Suppose that $M$ is a structure in $L_{\text{kernel}}$.
A surface substructure $C(s)$ for a surface $s \in M$ is the following set:

$$C(s) = \{ e : e \in M, (e, s) \in \text{part} \}$$

Intuitively, a surface substructure is the set of scene objects which are part of the surface. In terms of the incidence structures, the surface substructure for an element is the set of scene elements in the flag of the element. In Figure 3.2, the surface substructure for $s_1$ is $\{e_1, e_2, e_3, e_4, p_1, p_2, p_3, p_4\}$, and the surface substructure for $s_2$ is $\{e_5, e_6, e_7, p_5, p_6, p_7\}$.

We will now define the class of structures which will formalize the intended interpretations we have for the relations in $L_{\text{kernel}}$. In particular, for each relation, we will specify some property of the structures such that any structure with this property is isomorphic to the extension of the relation. There will also be constraints governing which structures can possibly be common substructures of the same structure in the class. The objective of this definition will be to show that any structure in this class will correspond to our intended intuitive interpretations, and also be a model of the axioms which we will present in our axiomatic theory. In addition, we will later show that any model of the axiomatic theory is isomorphic to some structure in this class.

To define the class of structures, we specify conditions which each structure in the class must satisfy, as well as conditions on the class as a whole. In the following definition of the class $M^*_{\text{kernel}}$, conditions 1-3 characterize the extensions of the relations denoted by the unary image and scene predicates in $L_{\text{kernel}}$. Condition 4 characterizes the extension of the relation denoted by the binary image predicate.
in and condition 5 characterizes the extension of the relation denoted by the binary scene predicate part. Condition 6 characterizes the extension of the relation denoted by the depiction predicate \( \Delta \). Properties 7, 8, and 9 characterize how the various substructures previously defined can be combined together to construct a structure in the class \( M^\text{kernel}_t \). Finally, condition 10 constrains the class of structures as a whole, by guaranteeing that every structure is an extension of some set of minimal structures in the class.

**Definition 3.5** Let \( M^\text{kernel}_t \) be the following class of structures in \( \mathcal{L}^\text{kernel} \), such that for any \( M \in M^\text{kernel}_t \), we have

1. Elements of \( M \) are divided into two disjoint sets:

   \[ M = I \cup C \]

   such that

   \[ I = \{ i : (i) \in \text{image}\_\text{object} \} \]

   \[ C = \{ c : (c) \in \text{scene}\_\text{object} \} \]

   Elements of \( I \) will be referred to as image objects and elements of \( C \) will be referred to as scene objects.

2. Elements of \( I \) are divided into three pairwise disjoint sets

   \[ I = Q \cup L \cup R \]

   where

   \[ Q = \{ q : (q) \in \text{pixel} \} \]

   \[ L = \{ l : (l) \in \text{line} \} \]

   \[ R = \{ r : (r) \in \text{region} \} \]

3. Elements of \( C \) are divided into three pairwise disjoint sets

   \[ C = P \cup E \cup S \]

   where

   \[ P = \{ p : (p) \in \text{point} \} \]

   \[ E = \{ e : (e) \in \text{edge} \} \]

   \[ S = \{ s : (s) \in \text{surface} \} \]

4. Elements of \( I \) form tripartite incidence structures without rank, where \( \text{in} \) is the transitive incidence relation:

   \[ \mathcal{I} = \langle I, \text{in} \rangle \]

   with the following additional conditions:

   (a) Within \( \mathcal{I} \), elements of \( Q \) have type 1, elements of \( L \) have type 2, and elements of \( R \) have type 3. Intuitively, regions contain lines and lines contain pixels.

   (b) Every element in \( R \) must be incident with at least three elements of \( L \), i.e., every region must contain at least three lines.

   (c) Every element in \( L \) must be incident with at least two elements of \( Q \), i.e., every line must contain at least two pixels.
(d) An element in \( L \) can be incident with at most two elements of \( R \), i.e., a line can be in at most two regions.

5. Elements of \( C \) form tripartite incidence structures of rank 3, where \( \text{part} \) is the transitive incidence relation:

\[
C = (C, \text{part})
\]

with the following additional conditions:

(a) Within \( C \), elements of \( P \) have type 1, elements of \( E \) have type 2, and elements of \( S \) have type 3. Intuitively, surfaces contain edges and edges contain points.

(b) Every element in \( S \) must be incident with at least three elements of \( E \), i.e., every surface must contain at least three edges.

(c) Every element in \( E \) must be incident with at least two elements of \( P \), i.e., every edge must contain at least two points.

(d) An element in \( P \) can be incident with at most two elements of \( E \), i.e., a point can be in at most two edges.

(e) \( C \) is the union of the surface substructures for each surface \( s \in M \).

6. The extension of \( \Delta \) is isomorphic to a directed bipartite graph \( G = (I, C, D) \) where \( I \) is the set of image objects, \( C \) is the set of scene objects and the set \( D \) of arcs in \( G \) must satisfy the following conditions:

(a) If \((i, c) \in D\) then

\[
(i, c) \in Q \times P
\]

or

\[
(i, c) \in L \times E
\]

or

\[
(i, c) \in R \times S
\]

(b) If \( e \in E \) and \( l \in L \) and

\[
(l, e) \in D
\]

then there exist elements \( p \in P \) and \( q \in Q \) such that

\[
(p, e) \in \text{part},
\]

\[
(q, l) \in \text{in},
\]

and

\[
(q, p) \in D
\]

In other words, if an edge is depicted by a line, then there exists a point that is part of the edge and a pixel contained in the line such that the pixel depicts the point.

(c) If \( e \in E \) and \( l \in L \) and

\[
(l, e) \in D
\]

and \((e, s) \in \text{part}, \) and there exist distinct regions \( r_1, r_2 \) such that \((l, r_1) \in \text{in} \) and \((l, r_2) \in \text{in} \), then either

\[
(r_1, s) \in D \text{ or } (r_2, s) \in D.
\]

In other words, if a line is in two distinct regions, and the line depicts an edge, then the surface containing the edge is depicted by one of the two regions.

7. \( C = C_{\text{dep}} \cup C_{\text{nondep}} \)

where
Chapter 3. Images, Scenes, and Depiction

- $C_{dep}$ is the substructure of $C$ consisting of surface substructures such that for each surface substructure $C(s_i)$ there exists an element $c \in C(s_i)$ and an element $i \in I$ such that
  \[(i, c) \in \Delta\]

  Intuitively, this is the incidence structure consisting of all surfaces which are either depicted or which contain a depicted part.

- $C_{nondep}$ is the substructure of $C$ consisting of surface substructures such that for all surface substructures $C(s_i)$ there do not exist elements $c \in C(s_i)$ and $i \in I$ such that
  \[(i, c) \in \Delta\]

  Intuitively, this is the incidence structure consisting of all surfaces which are neither depicted nor contain any depicted parts.

8. $I = I_{dep} \cup I_{nondep}$ where

- $I_{dep}$ is the substructure of $I$ consisting of image substructures such that for each image substructure $I(i_j)$ there exists an element $i \in I(i_j)$ and an element $c \in C$ such that
  \[(i, c) \in \Delta\]

  Intuitively, this is the incidence structure consisting of all image objects which either depict a scene object, or which contain an element that depicts a scene object, or which are contained in an element that depicts a scene object.

- $I_{nondep}$ is the substructure of $I$ consisting of image substructures such that for each image substructure $I(i_j)$ there do not exist elements $i \in I(i_j)$ and $c \in C$ such that
  \[(i, c) \in \Delta\]

  Intuitively, this is the incidence structure consisting of all image objects which do not depict scene objects and do not contain any elements which depict scene objects.

9. Let $G_1$ be the subgraph of the graph isomorphic to $\Delta$ consisting of elements in $I_{dep} \cup C_{dep}$.

\[M = G_1 \cup I_{nondep} \cup C_{nondep}\]

10. Define a minimally depicted structure to be one in which $C = C_{dep}$.

  Define a minimally depicting structure to be one in which $I = I_{dep}$.

  Given these definitions, every structure $M \in M_1^{kernel}$ is an extension of a structure which is both a minimally depicted structure and a minimally depicting structure.

**Theorem 3.1** Existence Theorem for $M_1^{kernel}$

The class of structures $M_1^{kernel}$ exists and is nonempty.

**Proof:** Structures in $M_1^{kernel}$ are the union of the following classes of structures:

1. tripartite incidence structures without rank $(\mathcal{I})$
2. tripartite incidence structures with rank 3 $(\mathcal{C})$
3. directed bipartite graphs (extension of $\Delta$)
4. $C_{dep}, C_{nondep}$
5. $I_{dep}, I_{nondep}$
The existence of incidence structures follows from [Dembowski 68], and the existence of directed bipartite graphs is trivial. The existence of $C_{\text{dep}}, C_{\text{nondep}}; I_{\text{dep}}$, and $I_{\text{nondep}}$ follows easily from the existence of the directed bipartite graph. $C_{\text{nondep}}$ is the set of isolated scene objects in the directed bipartite graph for $\Delta$. and $I_{\text{nondep}}$ is the set of isolated image objects in the graph. $C_{\text{dep}}$ and $I_{\text{dep}}$ are the remaining sets of elements in the graph.

Since each of the classes of substructures exist, their union exists. □

We now present the axioms of $T_{\text{kernel}} \cup T_{\text{scene}}$, and show that the structures in $M^1_{\text{kernel}}$ are equivalent to the models of $T_{\text{kernel}} \cup T_{\text{scene}}$.

### 3.2 The Axioms in the Kernel Theory

We will now present the axioms for scene objects and the depiction relation between the image and the scene. We call this set of axioms the kernel theory $T_{\text{kernel}}$, since it axiomatizes the basic ontology of CardWorld, and it is the smallest set of axioms we need to prove the characterization theorems later in the chapter. We will actually split these axioms into two sets. $T_{\text{scene}}$ (Figures 3.7 and 3.8) axiomatizes only the scene objects and the part relation over scene objects. $T_{\text{kernel}} = T_{\text{image}} \cup T_{\text{kernel}}^\Delta$ will be the set of axioms in Figure 3.5, Figure 3.6, and Figure 3.9. $T_{\text{kernel}}^\Delta$ axiomatizes image objects and the in relation over image objects, while $T_{\text{kernel}}^\Delta$ axiomatizes the basic constraints on depiction.

#### 3.2.1 Image Axioms

Figure 3.5 and Figure 3.6 present the axioms for image objects and image relations. There are three disjoint types of image objects, as defined by axioms 3.1 - 3.4 of $T_{\text{image}}$ in Figure 3.5 – pixels, lines, and regions. A composition hierarchy is defined over image objects with the relation $\text{in}(x, y)$ with the intended interpretation that image object $x$ is contained in image object $y$. Thus, regions contain lines, lines contain pixels, and pixels do not contain any other image objects. This relation defined in axioms 3.5 - 3.12 of $T_{\text{image}}$ in Figure 3.5.

There are also several coherence axioms which place additional constraints on the in relation – every region must contain lines, every line must contain pixels, and a line can be in at most two regions. These constraints are specified in axioms 3.13 - 3.15 of $T_{\text{image}}$ in Figure 3.6.

#### 3.2.2 Scene Axioms

Figure 3.7 and Figure 3.8 present the axioms defined over scene objects and scene relations. There are three disjoint types of scene objects, as defined by axioms 3.16 - 3.19 of $T_{\text{scene}}$ in Figure 3.7 – points, edges, and surfaces. A composition hierarchy is defined over scene objects with the relation $\text{part}(x, y)$ with the intended interpretation that scene object $x$ is part of scene object $y$. Thus, points are part of edges, edges are part of surfaces, surfaces are not part of any scene object, and no scene object is part of a point. These relations are defined in axioms 3.20 - 3.27 of $T_{\text{scene}}$ in Figure 3.7.
All image objects are either regions, lines, or pixels.

\[(\forall x) \text{image} \_ \text{object}(x) \equiv \text{region}(x) \lor \text{line}(x) \lor \text{pixel}(x)\]  

(3.1)

The sets of regions, lines, and pixels are disjoint.

\[\forall x \neg (\text{pixel}(x) \land \text{line}(x))\]  

(3.2)

\[\forall x \neg (\text{pixel}(x) \land \text{region}(x))\]  

(3.3)

\[\forall x \neg (\text{line}(x) \land \text{region}(x))\]  

(3.4)

The in relation holds only over image objects.

\[(\forall x, y) \text{in}(x, y) \supset \text{image} \_ \text{object}(x) \land \text{image} \_ \text{object}(y)\]  

(3.5)

A region cannot be contained in another region.

\[\forall x, r) \text{in}(x, r) \land \text{region}(r) \supset \neg \text{region}(x)\]  

(3.6)

A region or line cannot be contained in a line.

\[\forall x, l) \text{in}(x, l) \land \text{line}(l) \supset \neg \text{region}(x) \land \neg \text{line}(x)\]  

(3.7)

No image object can be contained in a pixel.

\[\forall x, q) \text{in}(x, q) \land \text{pixel}(q) \supset \neg \text{region}(x) \land \neg \text{line}(x) \land \neg \text{pixel}(x)\]  

(3.8)

A region cannot be contained in any other image object.

\[\forall x, r) \text{in}(r, x) \land \text{region}(r) \supset \neg \text{region}(x) \land \neg \text{line}(x) \land \neg \text{pixel}(x)\]  

(3.9)

A line cannot be contained in another line or in a pixel.

\[\forall x, l) \text{in}(l, x) \land \text{line}(l) \supset \neg \text{line}(x) \land \neg \text{pixel}(x)\]  

(3.10)

A pixel cannot be contained in another pixel.

\[\forall x, q) \text{in}(q, x) \land \text{pixel}(q) \supset \neg \text{pixel}(x)\]  

(3.11)

The in relation is transitive.

\[\forall l, r, v) \text{line}(l) \land \text{region}(r) \land \text{in}(l, r) \land \text{in}(v, l) \supset \text{in}(v, r)\]  

(3.12)

Figure 3.5: $T_{image}^k$: Image Axioms in $T_{kernel}$
Every region contains at least three distinct lines.

\[(\forall r) \text{region}(r) \supset \]
\[(\exists l_1, l_2, l_3) \text{line}(l_1) \land \text{line}(l_2) \land \text{line}(l_3) \land \text{in}(l_1, r) \land \text{in}(l_2, r) \land \text{in}(l_3, r) \land (l_1 \neq l_2) \land (l_1 \neq l_3) \land (l_2 \neq l_3) \quad (3.13)\]

Every line contains at least two distinct pixels.

\[(\forall l) \text{line}(l) \supset (\exists q_1, q_2) \text{pixel}(q_1) \land \text{pixel}(q_2) \land \text{in}(q_1, l) \land \text{in}(q_2, l) \land (q_1 \neq q_2) \quad (3.14)\]

A line is contained in at most two regions.

\[(\forall l, r_1, r_2, r_3) \text{line}(l) \land \text{region}(r_1) \land \text{region}(r_2) \land \text{region}(r_3) \land \text{in}(l, r_1) \land \text{in}(l, r_2) \land \text{in}(l, r_3)
\]
\[\supset (r_3 = r_1) \lor (r_2 = r_1) \quad (3.15)\]

Figure 3.6: $T_{\text{image}}$: Image Axioms in $T_{\text{kernel}}$ (cont.)

There are also several coherence axioms for scene objects – every edge is part of some surface, and every point is part of some edge. Moreover, edges and points are parts of unique surfaces. These relations are defined in axioms 3.28 - 3.33 of $T_{\text{scene}}$ in Figure 3.8.

Of particular interest is axiom 3.16 of $T_{\text{kernel}}$, which can be considered to be a scene object closure assumption. This axiom states that the only scene objects are points, edges, and surfaces. It thus eliminates the possibility that unknown scene objects exist. It is this axiom that restricts this thesis to shape-based object recognition; intuitively the shape of a surface is defined only by its component edges. This axiom eliminates the existence of such non-shape based scene objects such as surface markings, color changes, scratches, or textures.

### 3.2.3 Depiction Axioms

In addition to axioms defining the image and scene domains, we need axioms to define how scene objects are depicted by image objects. To do this we define the binary relation $\Delta(i, s)$ with the intended interpretation that image object $i$ depicts scene object $s$. The depiction axioms for scene objects are found in Figure 3.9. Thus points are depicted by pixels, edges are depicted by lines, and surfaces are depicted by regions. The following sentences are easy consequences of the depiction axioms, using axiom 3.16 of $T_{\text{scene}}$:

**Proposition 3.1**

\[T_{\text{kernel}} \models (\forall i, s) \text{pixel}(i) \land \Delta(i, s) \supset \text{point}(s)\]
\[T_{\text{kernel}} \models (\forall i, s) \text{line}(i) \land \Delta(i, s) \supset \text{edge}(s)\]
\[T_{\text{kernel}} \models (\forall i, s) \text{region}(i) \land \Delta(i, s) \supset \text{surface}(s)\]
All scene objects are either points, edges, or surfaces.

\[(\forall x) \text{scene}\_\text{object}(x) \equiv \text{point}(x) \lor \text{edge}(x) \lor \text{surface}(x)\]  \hspace{1cm} (3.16)

The sets of points, edges, and surfaces are disjoint.

\[\forall x \neg (\text{point}(x) \land \text{edge}(x))\]  \hspace{1cm} (3.17)

\[\forall x \neg (\text{point}(x) \land \text{surface}(x))\]  \hspace{1cm} (3.18)

\[\forall x \neg (\text{edge}(x) \land \text{surface}(x))\]  \hspace{1cm} (3.19)

The part relation is restricted to scene objects.

\[(\forall x, y) \text{part}(x, y) \supset \text{scene}\_\text{object}(x) \land \text{scene}\_\text{object}(y)\]  \hspace{1cm} (3.20)

A surface cannot be part of another surface.

\[(\forall x, s) \text{part}(x, s) \land \text{surface}(s) \supset \neg \text{surface}(x)\]  \hspace{1cm} (3.21)

A surface or an edge cannot be part of an edge.

\[(\forall x, e) \text{part}(x, e) \land \text{edge}(e) \supset \neg \text{surface}(x) \land \neg \text{edge}(x)\]  \hspace{1cm} (3.22)

Nothing can be part of a point.

\[(\forall x, p) \text{part}(x, p) \land \text{point}(p) \supset \neg \text{surface}(x) \land \neg \text{edge}(x) \land \neg \text{point}(x)\]  \hspace{1cm} (3.23)

A surface cannot be part of any other scene object.

\[(\forall x, s) \text{part}(s, x) \land \text{surface}(s) \supset \neg \text{surface}(x) \land \neg \text{edge}(x) \land \neg \text{point}(x)\]  \hspace{1cm} (3.24)

An edge cannot be part of an edge or a point.

\[(\forall x, e) \text{part}(e, x) \land \text{edge}(e) \supset \neg \text{edge}(x) \land \neg \text{point}(x)\]  \hspace{1cm} (3.25)

A point cannot be part of a point.

\[(\forall x, p) \text{part}(p, x) \land \text{point}(p) \supset \neg \text{point}(x)\]  \hspace{1cm} (3.26)

The part relation is transitive.

\[(\forall e, s, v) \text{edge}(e) \land \text{surface}(s) \land \text{part}(e, s) \land \text{part}(v, e) \supset \text{part}(v, s)\]  \hspace{1cm} (3.27)

---

**Figure 3.7:** $T_{\text{scene}}$: Scene axioms in $T_{\text{kernel}}$
Every edge is part of a surface.

\[ \forall x \ edge(x) \supset \exists s \ surface(s) \land part(x, s) \] (3.28)

Every point is part of an edge.

\[ \forall x \ point(x) \supset \exists e \ edge(e) \land part(x, e) \] (3.29)

A point is a part of a unique surface.

\[ (\forall v, s, s') \ part(v, s) \land part(v, s') \land point(v) \land surface(s) \land surface(s') \supset s = s' \] (3.30)

An edge is part of a unique surface.

\[ (\forall e, s, s') \ part(e, s) \land part(e, s') \land edge(e) \land surface(s) \land surface(s') \supset s = s' \] (3.31)

Every surface contains at least three edges.

\[ (\forall s) surface(s) \supset (\exists e_1, e_2, e_3) \ edge(e_1) \land edge(e_2) \land edge(e_3) \land (e_1 \neq e_2) \land (e_1 \neq e_3) \land (e_2 \neq e_3) \land part(e_1, s) \land part(e_2, s) \land part(e_3, s) \] (3.32)

Every edge contains at least two points.

\[ (\forall e) \ edge(e) \supset (\exists p_1, p_2) \ point(p_1) \land point(p_2) \land (p_1 \neq p_2) \land part(p_1, e) \land part(p_2, e) \] (3.33)

Figure 3.8: \( T_{\text{scene}} \): Scene axioms in \( T_{\text{kernel}} \) (cont.)
All elements are either scene objects or image objects.

\[ \forall x \text{scene.object}(x) \equiv \neg \text{image.object}(x) \]  
(3.34)

Depiction is a relation between image objects and scene objects.

\[ (\forall i, s) \Delta(i, s) \supset \text{image.object}(i) \land \text{scene.object}(s) \]  
(3.35)

Points are depicted by pixels.

\[ (\forall i, s) \text{point}(s) \land \Delta(i, s) \supset \text{pixel}(i) \]  
(3.36)

Edges are depicted by lines.

\[ (\forall i, s) \text{edge}(s) \land \Delta(i, s) \supset \text{line}(i) \]  
(3.37)

Surfaces are depicted by regions.

\[ (\forall i, s) \text{surface}(s) \land \Delta(i, s) \supset \text{region}(i) \]  
(3.38)

Lines do not depict points or surfaces.

\[ (\forall i, s) \text{line}(i) \land \Delta(i, s) \supset \neg \text{point}(s) \land \neg \text{surface}(s) \]  
(3.39)

Pixels do not depict edges or surfaces.

\[ (\forall i, s) \text{pixel}(i) \land \Delta(i, s) \supset \neg \text{edge}(s) \land \neg \text{surface}(s) \]  
(3.40)

Regions do not depict points or edges.

\[ (\forall i, s) \text{region}(s) \land \Delta(i, s) \supset \neg \text{point}(s) \land \neg \text{edge}(s) \]  
(3.41)

If an edge is depicted, then there exists a point contained in the edge which is depicted by a pixel in the line depicting the edge.

\[ (\forall l, e) \Delta(l, e) \supset (\exists p, q) \text{part}(p, e) \land \text{in}(q, l) \land \Delta(q, p) \]  
(3.42)

If a line is in two regions, then the surface containing the edge depicted by the line is depicted by one of these two regions.

\[ (\forall l, e, s, r, r') \Delta(l, e) \land \text{part}(e, s) \land \text{in}(l, r) \land \text{in}(l, r') \land r \neq r' \supset (\Delta(r, s) \lor \Delta(r', s)) \]  
(3.43)

Figure 3.9: \(T_{\text{kernel}}^\Delta\): Depiction axioms in \(T_{\text{kernel}}\)
$T_{kernel}$ and $T_{scene}$ provide us with a basic set of axioms defining the objects that exist in images and scenes. We now proceed to prove several results that hold for these axioms. The most important theorem will be a characterization of the set of models of the theory.

### 3.3 Characterization Theorems for $T_{kernel}$

In the previous section we presented the axioms of $T_{kernel}$, but in what sense are these axioms a correct and complete formalization of images, scenes, and depiction? In this section we prove the characterization theorem for the axioms in $T_{kernel}$; this theorem will present an intuitive result for which the axioms in $T_{kernel}$ are necessary and sufficient conditions. The primary motivation for this approach is that the axioms should be defined by their semantic properties, not by their syntactic form. Thus our theory should be strong enough to prove the appropriate characterization theorem, yet also be the minimal theory that can prove this result.

#### 3.3.1 Satisfiability of $T_{kernel}$

In this section, we show that $T_{kernel} \cup T_{scene}$ is satisfiable, that is, there exists a structure in $L_{kernel}$ which is a model of the theory, and further, that the class of structures $M^i_{kernel}$ is equivalent to the set of models for the theory.

**Theorem 3.2** Any structure in $M^i_{kernel}$ is a model of $T_{kernel} \cup T_{scene}$.

**Proof:** Let $M$ be a structure in $M^i_{kernel}$.

By condition (1) in the definition of $M^i_{kernel}$, the elements of $M$ are divided into disjoint sets of image objects and scene objects. Thus,

$$M \models (\forall z) \text{scene-object}(z) \equiv \neg \text{image-object}(z)$$

**Image Axioms** By condition (2), image objects in $M$ are divided into three sets of regions, lines, and pixels. Thus,

$$M \models (\forall z) \text{image-object}(z) \equiv \text{region}(z) \lor \text{line}(z) \lor \text{pixel}(z)$$

These sets are disjoint, so that we also have

$$M \models \forall z (\neg \text{pixel}(z) \land \text{line}(z))$$

$$M \models \forall z (\neg \text{pixel}(z) \land \text{region}(z))$$

$$M \models \forall z (\neg \text{line}(z) \land \text{region}(z))$$

By condition (3), scene objects in $M$ are divided into three sets of surfaces, edges, and points. Thus,

$$M \models (\forall z) \text{scene-object}(z) \equiv \text{point}(z) \lor \text{edge}(z) \lor \text{surface}(z)$$

These sets are disjoint, so that we also have

$$M \models \forall z (\neg \text{point}(z) \land \text{edge}(z))$$

$$M \models \forall z (\neg \text{point}(z) \land \text{surface}(z))$$
\( M \models \forall x \neg(\text{edge}(x) \land \text{surface}(x)) \)

By condition (4), image elements in \( M \) form incidence structures without rank. Since regions are the only elements which have rank 3 in these structures, we have

\[
M \models (\forall x, r) (\text{in}(x, r) \land \text{region}(r)) \supset \neg\text{region}(x)
\]

\[
M \models (\forall x, l) (\text{in}(x, l) \land \text{line}(l)) \supset \neg\text{line}(x)
\]

\[
M \models (\forall x, q) (\text{in}(x, q) \land \text{pixel}(q)) \supset \neg\text{pixel}(x)
\]

Since lines are the only elements of rank 2 in these structures, we have

\[
M \models (\forall x, l) (\text{in}(x, l) \land \text{line}(l)) \supset \neg\text{line}(x)
\]

\[
M \models (\forall x, q) (\text{in}(x, q) \land \text{pixel}(q)) \supset \neg\text{pixel}(x)
\]

Since pixels are the only elements of rank 1 in these structures, we have

\[
M \models (\forall x, q) (\text{in}(x, q) \land \text{pixel}(q)) \supset \neg\text{pixel}(x)
\]

Since \text{in} is transitive, we have

\[
M \models (\forall l, r, v) (\text{line}(l) \land \text{region}(r) \land \text{in}(l, r) \land \text{in}(v, l)) \supset \text{in}(v, r)
\]

Next, we have

\[
M \models (\forall r) \text{region}(r) \supset (\exists l) (\text{line}(l) \land \text{in}(l, r))
\]

iff for any variable assignment \( \sigma \), then for each region \( \sigma(r) \in M \) there exists a line \( \sigma(l) \in M \) such that \( \langle \sigma(l), \sigma(r) \rangle \in \text{in} \). This follows from condition 4(b) in the definition of \( M^\text{kernel}_i \).

\[
M \models (\forall l) \text{line}(l) \supset (\exists q_1, q_2) (\text{pixel}(q_1) \land \text{pixel}(q_2) \land \text{in}(q_1, l) \land \text{in}(q_2, l) \land (q_1 \neq q_2))
\]

iff for any variable assignment \( \sigma \), then for each line \( \sigma(l) \in M \) there exists distinct pixels \( \sigma(q_1), \sigma(q_2) \in M \) such that

\[
\langle \sigma(q_1), \sigma(l) \rangle, \langle \sigma(q_2), \sigma(l) \rangle \in \text{in}
\]

This follows from condition 4(c) in the definition of \( M^\text{kernel}_i \).

\[
M \models (\forall l, r_1, r_2, r_3) \text{line}(l) \land \text{region}(r_1) \land \text{region}(r_2) \land \text{region}(r_3) \land \text{in}(l, r_1) \land \text{in}(l, r_2) \land \text{in}(l, r_3) \supset
\]

\[
(r_3 = r_1) \lor (r_2 = r_1)
\]

iff for any variable assignment \( \sigma \), then for each line \( \sigma(l) \in M \), there are at most two regions \( \sigma(r_1), \sigma(r_2) \in M \) such that

\[
\langle \sigma(l_1), \sigma(r) \rangle, \langle \sigma(l_2), \sigma(r) \rangle \in \text{in}
\]

This follows from condition 4(d) in the definition of \( M^\text{kernel}_i \).
Scene Axioms  By condition (5), scene elements $M$ form incidence structures with rank 3. Since surfaces are the only elements which have rank 3 in these structures, we have

$$M \models (\forall x, s) \text{part}(x, r) \land \text{surface}(s) \supset \neg \text{surface}(x)$$

$$M \models (\forall x, e) \text{part}(x, e) \land \text{edge}(e) \supset \neg \text{edge}(x)$$

$$M \models (\forall x, p) \text{part}(x, p) \land \text{point}(p) \supset \neg \text{point}(x)$$

Since edges are the only elements of rank 2 in these structures, we have

$$M \models (\forall x, e) \text{part}(x, e) \land \text{edge}(e) \supset \neg \text{edge}(x)$$

$$M \models (\forall x, p) \text{part}(x, p) \land \text{point}(p) \supset \neg \text{point}(x)$$

Since points are the only elements of rank 1 in these structures, we have

$$M \models (\forall x, p) \text{part}(x, p) \land \text{point}(p) \supset \neg \text{point}(x)$$

Since $\text{part}$ is transitive, we have

$$M \models (\forall e, s, v) \text{edge}(e) \land \text{surface}(s) \land \text{part}(e, s) \supset \text{part}(v, e)$$

Next, we have

$$M \models (\forall s) \text{surface}(s) \supset (\exists e_1, e_2, e_3) \text{edge}(e_1) \land \text{edge}(e_2) \land \text{edge}(e_3) \land (e_1 \neq e_2) \land (e_1 \neq e_3) \land (e_2 \neq e_3)$$

$$\land \text{part}(e_1, s) \land \text{part}(e_2, s) \land \text{part}(e_3, s)$$

iff for any variable assignment $\sigma$, then for each surface $\sigma(s) \in M$ there exist three distinct edges $\sigma(e_1), \sigma(e_2), \sigma(e_3) \in M$ such that

$$(\sigma(e_1), \sigma(s)), (\sigma(e_2), \sigma(s)), (\sigma(e_3), \sigma(s)), \in \text{part}$$

This follows from condition 5(b) in the definition of $M^\text{kernel}$.

$$M \models (\forall e) \text{edge}(e) \supset (\exists p_1, p_2) \text{point}(p_1) \land \text{point}(p_2) \land (p_1 \neq p_2) \land \text{part}(p_1, e) \land \text{part}(p_2, e)$$

iff for any variable assignment $\sigma$, then for each edge $\sigma(e) \in M$ there exists distinct points $\sigma(p_1), \sigma(p_2) \in M$ such that

$$(\sigma(p_1), \sigma(e)), (\sigma(p_2), \sigma(e)) \in \text{part}$$

This follows from condition 5(c) in the definition of $M^\text{kernel}$.

Depiction Axioms  By condition (6), the extension of $\Delta$ is isomorphic to a directed bipartite graph, so we have

$$M \models (\forall i, s) \Delta(i, s) \supset \text{image.object}(i) \land \text{scene.object}(s)$$

By condition (6a), we have

$$M \models (\forall i, s) \text{line}(i) \land \Delta(i, s) \supset \neg \text{point}(s) \land \neg \text{surface}(s)$$

$$M \models (\forall i, s) \text{pixel}(i) \land \Delta(i, s) \supset \neg \text{edge}(s) \land \neg \text{surface}(s)$$

$$M \models (\forall i, s) \text{region}(i) \land \Delta(i, s) \supset \neg \text{point}(s) \land \neg \text{edge}(s)$$

By condition (6b), we have

$$M \models (\forall i, e) \Delta(l, e) \supset (\exists p, q) \text{part}(p, e) \land \text{in}(q, l) \land \Delta(q, p)$$

By condition (6c), we have

$$M \models (\forall l, e, s, r, r') \Delta(l, e) \land \text{part}(e, s) \land \text{in}(l, r) \land \text{in}(l, r') \land r \neq r' \supset (\Delta(r, s) \lor \Delta(r', s))$$

By properties (7), (8), and (9), for any scene object $s$ and image object $i$, there are three cases:
Suppose \( s \) is either disconnected in the depiction graph (case 2), or it is connected to an image element \( i \) which is a pixel (case 1). Similarly \( i \) is either disconnected in the depiction graph (case 3), or it is connected to an scene element which is a point (case 1).

By condition (10) in the definition of \( \mathcal{M}_{kernel} \), \( \mathcal{M} \) is either a minimally depicting and minimally depicted structure (so that all elements satisfy case 1) or it is an extension of such a structure (so that there are elements which satisfy case 2 or 3).

Thus, if \( (i, s) \in \Delta \) then

\[
\langle i, s \rangle \in \text{pixel}
\]

so that

\[
\mathcal{M} \models (\forall i, s) \text{ point}(s) \land \Delta(i, s) \supset \text{pixel}(i)
\]

Similar arguments hold for the other classes of scene and image objects, so that we have:

\[
\mathcal{M} \models (\forall i, s) \text{ edge}(s) \land \Delta(i, s) \supset \text{line}(i)
\]

\[
\mathcal{M} \models (\forall i, s) \text{ surface}(s) \land \Delta(i, s) \supset \text{region}(i)
\]

Therefore, any structure \( \mathcal{M} \in \mathcal{M}_{kernel} \) satisfies every axiom in \( T_{kernel} \). □

### 3.3.2 Depicted Kernel Theorem

The structure theorem for the axioms in \( T_{kernel} \) is based on the notion of a parsimonious representation of the set of models for a theory. This idea has been used in the model theoretic discipline of stability theory ([Baldwin 88], [Hodges 93]) which attempts to find structure theorems for various theories – given the set of models for a theory, is there some way of representing this set without enumerating all models in the set? The central theorem of this section will represent the set of models of an image by those models that have minimal domains.

We first need to characterize the relationship among structures in the class.

**Definition 3.6** A structure \( \mathcal{M} \) is a substructure of a structure \( \mathcal{N} \) (denoted by \( \mathcal{M} < \mathcal{N} \)) if the domain of \( \mathcal{M} \) is a subset of the domain of \( \mathcal{N} \) and:

- Each \( n \)-placed relation \( R' \) of \( \mathcal{M} \) is the restriction to the domain of \( \mathcal{M} \) of the corresponding relation \( R \) of \( \mathcal{N} \).

- Each \( m \)-placed function \( G' \) of \( \mathcal{M} \) is the restriction to the domain of \( \mathcal{M} \) of the corresponding function \( G \) of \( \mathcal{N} \).

- Each constant of \( \mathcal{M} \) is the corresponding constant of \( \mathcal{N} \).

\( \mathcal{N} \) is an extension of \( \mathcal{M} \). A structure is minimal if it has no proper substructure.

**Definition 3.7** Two structures, \( \mathcal{M}, \mathcal{N} \) agree on a sentence \( \phi \) when \( \mathcal{M} \models \phi \) iff \( \mathcal{N} \models \phi \).
Figure 3.10: Examples of structures for the depicted kernel.
Example: Consider the image $I$ in Figure 3.10a.

Consider the following four structures in $M_i^{kernel}$, which all share the same domain of lines and regions:

$$\{r_1, r_2, r_3, r_4, l_1, l_2, l_3\}$$

- The domain of edges and surfaces in $M_1$ includes the set $\{e_1, e_2, e_3, s_1, s_3, s_4\}$,

$$\langle r_1, s_1 \rangle, \langle r_3, s_2 \rangle, \langle r_4, s_3 \rangle \in \Delta$$

$$\langle l_1, e_1 \rangle, \langle l_2, e_2 \rangle, \langle l_3, e_3 \rangle \in \Delta$$

$$\langle e_1, s_1 \rangle, \langle e_2, s_3 \rangle, \langle e_3, s_4 \rangle \in \text{part}$$

In this structure, each region except $r_2$ depicts a unique surface, and all surfaces are depicted by some region. The region $r_2$ does not depict a surface, but is merely background.

For this structure, we have

$$C_{dep} = \{s_1, e_1, s_3, e_2, s_4, e_3\}$$

$$C_{nondep} = \emptyset$$

$$I_{nondep} = \{r_2\}$$

- The domain of edges and surfaces in $M_2$ includes the set $\{e_1, e_2, e_3, s_1, s_2, s_3, s_4\}$,

$$\langle r_1, s_1 \rangle, \langle r_2, s_2 \rangle, \langle r_3, s_3 \rangle, \langle r_4, s_4 \rangle \in \Delta$$

$$\langle l_1, e_1 \rangle, \langle l_2, e_2 \rangle, \langle l_3, e_3 \rangle \in \Delta$$

$$\langle e_1, s_1 \rangle, \langle e_2, s_3 \rangle, \langle e_3, s_4 \rangle \in \text{part}$$

In this structure, the region $r_2$ depicts a surface in which all of its edges are intuitively occluded\(^3\).

For this structure, we have

$$C_{dep} = \{s_1, e_1, s_2, s_3, e_2, s_4, e_3\}$$

$$C_{nondep} = \emptyset$$

\(^3\)We rely for now on the intuition of occlusion; the following chapter will formalize this intuition.
- The domain of edges and surfaces in $\mathcal{M}_3$ includes the set \{e_1, e_2, e_3, s_1, s_2, s_3, s_4\}.

\[
\langle r_1, s_1 \rangle, \langle r_2, s_2 \rangle, \langle r_3, s_3 \rangle, \langle r_4, s_4 \rangle \in \Delta
\]

\[
\langle l_1, e_1 \rangle, \langle l_1, e_4 \rangle, \langle l_2, e_2 \rangle, \langle l_3, e_3 \rangle \in \Delta
\]

\[
\langle e_1, s_1 \rangle, \langle e_4, s_2 \rangle, \langle e_2, s_3 \rangle, \langle e_3, s_4 \rangle \in \text{part}
\]

In this structure, the region $r_2$ depicts a surface and the line $l_1$ depicts an edge in this surface: all other edges in the surface are not depicted; the edges in the surface $s_2$ are displayed in Figure 3.10(b).

For this structure, we have

\[C_{dep} = \{s_1, e_1, e_2, s_3, e_2, s_4, e_3\}\]

\[C_{nondep} = \emptyset\]

- The domain of edges and surfaces in $\mathcal{M}_4$ includes the set \{e_1, e_2, e_3, s_1, s_2, s_3, s_4, s_5\}.

\[
\langle r_1, s_1 \rangle, \langle r_2, s_2 \rangle, \langle r_3, s_3 \rangle, \langle r_4, s_4 \rangle \in \Delta
\]

\[
\langle l_1, e_1 \rangle, \langle l_2, e_2 \rangle, \langle l_3, e_3 \rangle \in \Delta
\]

\[
\langle e_1, s_1 \rangle, \langle e_2, s_3 \rangle, \langle e_3, s_4 \rangle \in \text{part}
\]

In this structure, we have a surface $s_5$ that is completely occluded by the surfaces in the scene.

For this structure, we have

\[C_{dep} = \{s_1, e_1, s_2, s_3, e_2, s_4, e_3\}\]

\[C_{nondep} = \{s_5\}\]

Given this set of structures, we have the following ordering over the set:

\[\mathcal{M}_1 < \mathcal{M}_2 < \mathcal{M}_4\]

\[\mathcal{M}_1 < \mathcal{M}_3\]

Note that $\mathcal{M}_3$ is not related to either $\mathcal{M}_2$ or $\mathcal{M}_4$, and that $\mathcal{M}_1$ is a minimal model. Also note that $\mathcal{M}_2$ and $\mathcal{M}_4$ agree on all depiction literals. \[\square\]
Thus, we can order the set of structures in $M^\text{Kernel}_k$ with respect to the set of depicted surfaces within the domain of a structure. Our next task is to characterize the minimal structures in the ordering.

**Definition 3.8** A **depicted kernel model** is a model of $T_{\text{Kernel}} \cup T_{\text{Scene}}$ which satisfies the Depicted Kernel Axiom (DKA):

\[
(\forall z) \text{scene}_z \cdot \text{object}(z) \supset \\
((\exists y) \Delta(y, z) \\
\forall (\exists x', y', z) \text{part}(x, z) \land \text{part}(x', z) \land \Delta(y', x') \\
\forall (\exists z, w) \text{part}(x, z) \land \Delta(w, z) \\
\forall (\exists z, w) \text{part}(z, x) \land \Delta(w, z))
\]

In other words, in depicted kernel models all scene objects are in some way “grounded” by depicted scene objects. They are either depicted, or they are part of a depicted scene object, or they contain a depicted scene object, or another part of the same scene object is depicted. This is a very strong property: we will later use closure assumptions over the image domain, and together with DKA, we can eliminate surfaces without names, since all scene objects can be given a name depending on the depicting image object.

In the preceding examples, we see that $M_4$ is not in the set of depicted kernel models of the image, since the surface $s_5$ is not depicted and has no depicted edges or points. $M_1$, $M_2$ and $M_3$ are in the set of depicted kernel models since all surfaces in their domains are depicted. In $M_2$, there are undepicted edges in the surface; however, they satisfy $\text{DKA}$ since they are part of the depicted surface $s_2$. Similarly, $M_3$ satisfies $\text{DKA}$, since there is an edge in the surface that is depicted (thus satisfying the second disjunct in the definition) and the surface containing the edge is depicted (satisfying the third disjunct of the definition.)

**Example:** Consider the image $I$ in Figure 3.10(c). Note that $\mathcal{L}(I) = \{l_1, l_2, r_1, r_2\}$

Consider two models $M_1, M_2$ of $T_{\text{Kernel}} \cup T_{\text{Scene}} \cup I$ with domains $M_1 = \{e_1, e_2, s_1, s_2, s_3\}$ and $M_2 = \{e_1, e_2, s_1, s_2, s_3\}$ such that within $M_1$ we have

\[
(r_1, s_1), (l_1, e_1), (l_2, e_2) \in \Delta
\]

\[
(e_1, s_2), (e_2, s_2) \in \text{part}
\]

and within $M_2$ we have

\[
(r_1, s_1), (l_1, e_1), (l_2, e_2) \in \Delta
\]

\[
(e_1, s_2), (e_2, s_3) \in \text{part}
\]

The scene in $M_1$ is displayed in Figure 3.10(d), and the scene in $M_2$ is displayed in Figure 3.10(e). Both models are kernel models of the image; although there exist undepicted surfaces in each model, they have depicted edges, thus satisfying the fourth disjunct in the definition of $\text{DKA}$. □
We can relativize the above definition to surfaces in the following way:

**Proposition 3.2**

\[ T_{\text{kernel}} \cup T_{\text{scene}} \cup DK \models (\forall s) \text{surface}(s) \supset ([\exists r] \Delta(r, s) \lor (\exists l, e) \text{part}(e, s) \land \Delta(l, e)) \]

Thus any model of \( T_{\text{kernel}} \cup T_{\text{scene}} \cup DK \) must contain only depicted surfaces: there can be no surfaces hiding behind other surfaces (that is, completely occluded), and there are no surfaces whose edges are completely undepicted due to errors in edge detection.

Before proceeding to the characterization theorem, we present the following property of the kernel axioms which will play a key role in the theorems of this section. It shows that all scene objects are either surfaces or are parts of surfaces.

**Lemma 3.1**

\[ T_{\text{scene}} \models (\forall z) \text{scene.object}(x) \equiv \text{surface}(x) \lor (\exists y) \text{part}(x, y) \land \text{surface}(y) \]

**Proof:** By axiom 3.16 of \( T_{\text{scene}} \) we have

\[ T_{\text{scene}} \models (\forall z) \text{scene.object}(x) \equiv \text{surface}(x) \lor \text{edge}(x) \lor \text{point}(x) \]

By axioms 3.25 and 3.29 we have

\[ T_{\text{scene}} \models (\forall z) \text{point}(x) \supset (\exists s) \text{surface}(s) \land \text{part}(x, s) \]

while axiom 3.28 gives

\[ T_{\text{scene}} \models (\forall z) \text{edge}(x) \supset (\exists s) \text{surface}(s) \land \text{part}(x, s) \]

By axiom 3.21 we have

\[ T_{\text{scene}} \models (\forall z, s) \text{part}(x, s) \land \text{surface}(s) \supset \text{point}(x) \lor \text{edge}(x) \]

so that we have

\[ T_{\text{scene}} \models (\forall z) (\text{edge}(x) \lor \text{point}(x)) \equiv (\exists s) \text{surface}(s) \land \text{part}(x, s) \]

\( \square \)

We will now see that the depicted kernel assumption characterizes the set of models for \( T_{\text{kernel}} \cup T_{\text{scene}} \):

**Theorem 3.3 Depicted Kernel Theorem**

Let \( \{M_1, \ldots\} \) be a set of models of \( T_{\text{kernel}} \) which agree on the extensions of all image relations. Any model \( M \) in this set is an extension of a unique depicted kernel model \( M' \) constructed by adjoining a set \( \{s_1, \ldots\} \) of undepicted surfaces to \( M' \).

**Proof:** Let \( M \) be a model which is not a depicted kernel model. Then the following set is nonempty:

\[ \Sigma = \{ x : M \models \text{scene.object}(x) \land \neg(\exists y) \Delta(y, x) \\
\land \neg(\exists y', z) \text{part}(x, z) \land \text{part}(x', z) \land \Delta(y', x') \\
\land \neg(\exists w) \text{part}(x, z) \land \Delta(w, z) \} \]
\[ \land \neg (\exists z, w) \text{part}(z, x) \land \Delta(w, z) \}

Note that all scene objects in \( \Sigma \) are undepicted, and that by Lemma 3.1, all of these undepicted scene objects are either surfaces or parts of surfaces.

Let \( \mathcal{M}' \) be the substructure of \( \mathcal{M} \) such that \( \mathcal{M}' = \mathcal{M} - \Sigma \).

Claim: \( \mathcal{M}' \) is a depicted kernel model of \( T_{\text{kernel}} \cup T_{\text{scene}} \).

Clearly, \( \mathcal{M}' \not \models DKA \), since all scene objects that violated \( DKA \) are in \( \Sigma \).

\( \mathcal{M} \) and \( \mathcal{M}' \) agree on the extension of \( \Delta \) since the only objects eliminated from the domain are undepicted scene objects.

Suppose \( \mathcal{M}' \) is not a model of \( T_{\text{kernel}} \cup T_{\text{scene}} \). Then \( \mathcal{M} \) and \( \mathcal{M}' \) must not agree on the extension of some scene relation in \( \mathcal{M}' \); however, this extension contains only elements in \( \mathcal{M}' \) and by the following lemma, \( \mathcal{M} \) and \( \mathcal{M}' \) agree on the extensions of relations with those objects.

**Lemma 3.2** For any variable assignment \( \sigma \), if

\[ \mathcal{M}, \sigma \models \text{part}(a, b) \]

then \( \sigma(a) \in \Sigma \) iff \( \sigma(b) \in \Sigma \).

**Proof:** Suppose \( \mathcal{M}, \sigma \models \text{part}(a, b) \) and \( \sigma(a) \in \Sigma \). By axiom 3.16 of \( T_{\text{scene}} \) there are three cases:

1. \( \mathcal{M}, \sigma \models \text{edge}(a) \).

   By axiom 3.16 of \( T_{\text{scene}} \) and axiom 3.25,

   \[ \mathcal{M}, \sigma \models \text{surface}(b) \]

   By the definition of \( \Sigma \) we have

   \[ \mathcal{M}, \sigma \models ((\forall w) \neg \Delta(w, b)) \land ((\forall z, y) \text{part}(z, b) \supset \neg \Delta(y, z)) \]

   so that \( \sigma(b) \in \Sigma \).

2. \( \mathcal{M}, \sigma \models \neg \text{surface}(a) \), by axiom 3.16 of \( T_{\text{scene}} \) and axiom 3.24.

3. \( \mathcal{M}, \sigma \models \text{point}(a) \).

   By axiom 3.16 of \( T_{\text{scene}} \) and axiom 3.26, we have either \( \mathcal{M}, \sigma \models \text{surface}(b) \) or \( \mathcal{M}, \sigma \models \text{edge}(b) \).

   By the definition of \( \Sigma \) we have

   \[ \mathcal{M}, \sigma \models ((\forall w) \neg \Delta(w, b) \land (\forall r, s) \text{part}(b, s) \supset \neg \Delta(r, s)) \]

   \[ \land (\forall e, s, r) \text{part}(b, s) \land \text{part}(e, s) \supset \neg \Delta(r, s)) \land (\forall p, q) \text{part}(p, b) \supset \neg \Delta(q, p)) \]

   so that \( \sigma(b) \in \Sigma \).

Now suppose \( \mathcal{M}, \sigma \models \text{part}(a, b) \) and \( \sigma(b) \in \Sigma \). Again, by axiom 3.16 of \( T_{\text{scene}} \) there are three cases:

1. \( \mathcal{M}, \sigma \models \text{edge}(b) \).

   By axiom 3.16 of \( T_{\text{scene}} \) and axiom 3.22, \( \mathcal{M}, \sigma \models \text{point}(a) \). By the definition of \( \Sigma \) we get

   \[ \mathcal{M}, \sigma \models ((\forall w) \neg \Delta(w, a)) \land (\forall z, y) \text{part}(a, z) \supset \neg \Delta(y, a)) \]

   so that \( \sigma(a) \in \Sigma \).

2. \( \mathcal{M}, \sigma \models \neg \text{point}(b) \) by SOCA and axiom 3.23.

3. \( \mathcal{M}, \sigma \models \text{surface}(b) \).

   By axiom 3.16 of \( T_{\text{scene}} \) and axiom 3.21, we have either \( \mathcal{M}, \sigma \models \text{point}(a) \) or \( \mathcal{M}, \sigma \models \text{edge}(a) \).

   By the definition of \( \Sigma \) we have

   \[ \mathcal{M}, \sigma \models ((\forall w) \neg \Delta(w, a) \land (\forall r, s) \text{part}(a, s) \supset \neg \Delta(r, s)) \]

   \[ \land (\forall e, s, r) \text{part}(a, s) \land \text{part}(e, s) \supset \neg \Delta(r, s)) \land (\forall p, q) \text{part}(p, a) \supset \neg \Delta(q, p)) \]

   so that \( \sigma(a) \in \Sigma \).
Thus if $\mathcal{M}'$ is not a model then neither is $\mathcal{M}$, which contradicts our original assumption.

We now proceed to the proof of the uniqueness of the characterization. Let $\mathcal{M}$ be a model of $T_{\text{kernel}} \cup T_{\text{scene}}$ and suppose that $\mathcal{M}$ is an extension of two depicted kernel models $\mathcal{M}_1, \mathcal{M}_2$. By the above argument, $\mathcal{M}, \mathcal{M}_1$ and $\mathcal{M}, \mathcal{M}_2$ must agree on the extensions of the depiction relation and all scene and image relations. $\mathcal{M}_1$ and $\mathcal{M}_2$ must therefore be isomorphic. □

We can consider $DKA$ to be an assumption that eliminates the class of anomalous models in which there exist surfaces that are totally undepicted.

**Corollary 3.1** All minimal models of $T_{\text{kernel}} \cup T_{\text{scene}}$ are depicted kernel models.

We will now explore the structure of non-minimal models of $T_{\text{kernel}} \cup T_{\text{scene}}$ through the following theorems.

**Theorem 3.4** Given two models $\mathcal{M}, \mathcal{N}$ of $T_{\text{kernel}} \cup T_{\text{scene}}$ which agree on the extensions of all image relations, $\mathcal{M} < \mathcal{N}$ iff there exists a surface $s$ such that $s \in \mathcal{N}, s \notin \mathcal{M}$ and for all scene objects $s_i$ such that $(s_i, s) \in \text{part}^N$, we have $s_i \in \mathcal{N}, s_i \notin \mathcal{M}$.

**Proof:** By definition, $\mathcal{M} < \mathcal{N}$ iff there exists an object in the domain of $\mathcal{N}$ that is not an object in the domain of $\mathcal{M}$. By the hypothesis of the theorem, $\mathcal{M}$ and $\mathcal{N}$ can differ only on the set of scene objects. By Lemma 3.1, every scene object is either a surface or is part of a surface. Thus there exists a surface $s$ such that $s \in \mathcal{N}$ and $s \notin \mathcal{M}$; we need to show that all scene objects that are a part of $s$ also satisfy this property. By axioms 3.30 and 3.31, points and edges are part of unique surfaces; if $s$ is not in the domain of $\mathcal{M}$, then neither can any scene object that is part of $s$. □

This theorem is an important auxiliary result to the Depicted Kernel Theorem, since it formalizes the intuition that models of $T_{\text{kernel}} \cup T_{\text{scene}}$ are "composed" of separate surfaces, and that surfaces are the "pieces" we use to construct models.

**Theorem 3.5** If $\mathcal{M}, \mathcal{N}$ are depicted kernel models of $T_{\text{kernel}} \cup T_{\text{scene}}$ which agree on the extensions of all image relations and $\mathcal{M} < \mathcal{N}$ then there exist scene objects $s, s'$ and image object $i$ such that $s' \in \mathcal{N}, s' \notin \mathcal{M}$, $i \in \mathcal{M}$, $i \notin \mathcal{N}$, and either

- $i$ depicts both $s$ and $s'$ in $\mathcal{N}$, but it only depicts $s$ in $\mathcal{M}$, or
- $i$ depicts $s'$ in $\mathcal{N}$ and $i$ does not depict any surface in $\mathcal{M}$.

**Proof:** By the previous theorem, there must exist a surface $s'$ such that $s' \in \mathcal{N}, s' \notin \mathcal{M}$. Since $\mathcal{M}, \mathcal{N}$ are depicted kernel models we have for any variable assignment $\sigma$,

$$\mathcal{N}, \sigma \models (\exists r) \Delta(r, s) \lor (\exists e) \text{part}(e, s) \land \Delta(l, e)$$

By the hypothesis of the theorem, these depicting image objects must be in both $\mathcal{M}$ and $\mathcal{N}$. There are two cases — either they depict scene objects in $\mathcal{M}$ (in which case they depict multiple scene objects in $\mathcal{N}$) or they do not depict scene objects in $\mathcal{M}$. □
Thus, non-minimal kernel models either have multiply depicting image objects or they have image objects that need not depict scene objects. For example, consider the models of the image in Figure 3.10(a): $M_2$ and $M_3$ are non-minimal kernel models. In the first model, $r_2$ depicts a surface, but in the minimal model it does not depict any scene object. In $M_3$, $l_1$ depicts multiple edges in the surfaces $s_1$ and $s_2$.

Thus there are two aspects to the Depicted Kernel Theorem that are reflected in the later theorems. First, all models are extensions of a unique kernel model, and second, nonkernel models can be constructed piecewise from kernel models by adding undepicted surfaces.

### 3.3.3 Depicting Kernel Theorem

Within the axioms of $T_{kernel} \cup T_{scene}$, there is a certain symmetry between image objects and scene objects. Both form incidence structures, and in both cases, we can identify substructures based on the depiction relation. We have just seen how models of $T_{kernel} \cup T_{scene}$ can be decomposed into depicted and undepicted scene objects. By this intuition of symmetry, we can expect an analogous theorem for depicting and nondepicting image objects.

**Definition 3.9** A depicting kernel model is a model of $T_{kernel} \cup T_{scene}$ which satisfies the Depicting Kernel Axiom (DKA):

$$(\forall x) \text{image.object}(x) \supset (\exists y) \Delta(x, y)$$

$$\lor (\exists w, z) \text{in}(w, x) \land \Delta(w, z)$$

In other words, in depicting kernel models all image objects are in some way related to depicted scene objects. All image objects either depict a scene object, or they contain an image object which depict a scene object.

Note that there may exist models in which regions in the image do not depict a surface, but rather are simply background (see Figure 3.10(a)). Such models will still be depicting kernel models, since the lines in the region do depict edges in other surfaces.

The analogous theorem to the Depicted Kernel Theorem is thus the following:

**Theorem 3.6** Depicting Kernel Theorem

Let $\{M_1, \ldots\}$ be a set of models of $T_{kernel} \cup T_{scene}$ which agree on the extensions of all scene relations. Any model $M$ in this set is an extension of a unique depicting kernel model $M'$ constructed by adjoining a set $\{i_1, \ldots\}$ of nondepicting image objects to $M'$.

**Proof:** Let $M$ be a model which is not a depicting kernel model. Then the following set is nonempty:

$$\Sigma = \{x : M \models \text{image.object}(x) \land \neg(\exists y) \Delta(x, y)$$

$$\land \neg(\exists w, z) \text{in}(w, x) \land \Delta(w, z)\}$$

Note that none of the image objects in $\Sigma$ depict scene objects.

Let $M'$ be the substructure of $M$ such that $M' = M - \Sigma$.

**Claim:** $M'$ is a depicting kernel model of $T_{kernel} \cup T_{scene}$.

Clearly, $M' \not\models$ DKA, since all image objects that violated DKA are in $\Sigma$. 

.\( M \) and \( M' \) agree on all the extension of \( \Delta^M \) since the only objects eliminated from the domain are nondepicting image objects.

Suppose \( M' \) is not a model of \( T_{kernel} \cup T_{scene} \). Then \( M \) and \( M' \) must not agree on the extension of some image relation in \( M' \); however, this extension contains only elements in \( M' \) and by the following lemma, \( M \) and \( M' \) agree on the extensions of relations with those objects.

**Lemma 3.3** For any variable assignment \( \sigma \), if
\[
M, \sigma \models in(a, b)
\]
then \( \sigma(a) \in \Sigma \) iff \( \sigma(b) \in \Sigma \).

**Proof:** Suppose \( M, \sigma \models in(a, b) \) and \( \sigma(a) \in \Sigma \). By axiom 3.1 of \( T_{image} \) there are three cases:

1. \( M, \sigma \models line(a) \).
   By axiom 3.1 of \( T_{image} \) and axiom 10,
   \[
   M, \sigma \models region(b)
   \]
   By the definition of \( \Sigma \) we have
   \[
   M, \sigma \models ((\forall w) \neg \Delta(b, w)) \land ((\forall z, y) \ (in(b, z) \supset \neg \Delta(z, y))
   \]
   so that \( \sigma(b) \in \Sigma \).

2. \( M, \sigma \models \neg region(a) \), by axiom 3.1 of \( T_{image} \) and axiom 9.

3. \( M, \sigma \models pixel(a) \).
   By axiom 3.1 of \( T_{image} \) and axiom 3.10, we have either \( M, \sigma \models region(b) \) or \( M, \sigma \models line(b) \).
   By the definition of \( \Sigma \) we have
   \[
   M, \sigma \models (\forall x) \neg \Delta(b, x) \land ((\forall z, r_2) \ in(r_2, b) \supset \neg \Delta(b, z))
   \]
   so that \( \sigma(b) \in \Sigma \).

Now suppose \( M, \sigma \models in(a, b) \) and \( \sigma(b) \in \Sigma \). Again, by axiom 3.1 of \( T_{image} \) there are three cases:

1. \( M, \sigma \models line(b) \)
   By axiom 3.1 of \( T_{image} \) and axiom 3.7, \( M, \sigma \models pixel(a) \). By the definition of \( \Sigma \) we get
   \[
   M, \sigma \models ((\forall w) \neg \Delta(a, w))
   \]
   so that \( \sigma(a) \in \Sigma \).

2. \( M, \sigma \models \neg pixel(b) \) by axiom 3.1 and 3.8.

3. \( M, \sigma \models region(b) \).
   By axiom 3.1 of \( T_{image} \) and axiom 3.6, we have either \( M, \sigma \models pixel(a) \) or \( M, \sigma \models line(a) \).
   By the definition of \( \Sigma \) we have
   \[
   M, \sigma \models (\forall x) \neg \Delta(a, x) \land ((\forall z, r_2) \ in(r_2, a) \supset \neg \Delta(a, z))
   \]
   so that \( \sigma(a) \in \Sigma \).

Thus if \( M' \) is not a model then neither is \( M \), which contradicts our original assumption.

We now proceed to the proof of the uniqueness of the characterization. Let \( M \) be a model of \( T_{kernel} \cup T_{scene} \) and suppose that \( M \) is an extension of two depicted kernel models \( M_1, M_2 \). By the above argument, \( M, M_1 \) and \( M, M_2 \) must agree on the extensions of the depiction relation and all scene and image relations. \( M_1 \) and \( M_2 \) must therefore be isomorphic.

\( \square \)
Another perspective on the Depicting Kernel Theorem is to consider the preservation of models under noise for some scene. This theorem tells us that if we have a model of $T_{\text{kernel}} \cup T_{\text{scene}}$ for some image and some scene, this model is preserved for all images which are extensions of the original image with added spurious noise, that is, image objects which do not depict any scene object.

As with the Depicted Kernel Theorem, we can characterize the structure of these non-minimal models of $T_{\text{kernel}} \cup T_{\text{scene}}$ through the following theorems.

**Theorem 3.7** Given two models $M, N$ of $T_{\text{kernel}} \cup T_{\text{scene}}$ which agree on the extensions of all scene relations, $M < N$ iff there exists an image object $i$ such that $i \in N, i \notin M$ and there exist image objects $i_j$ such that $(i_j, i) \in N, i_j \in N$, and $i_j \notin M$.

**Proof:** By definition, $M < N$ iff there exists an object in the domain of $N$ that is not an object in the domain of $M$. By the hypothesis of the theorem, $M$ and $N$ can differ only on the set of image objects. Thus there exists an image object $i$ such that $i \in N$ and $i \notin M$: further, $N$ does not satisfy IDKA, so that $i$ does not depict a scene object, and none of the image objects contained in $i$ depict a scene object. Since all of the image objects in $M$ depict a scene object, none of the nondepicting image objects in $N$ can also be in $M$.

\[\square\]

### 3.3.4 Characterizing Models of $T_{\text{kernel}}$

**Theorem 3.8** Any model of $T_{\text{kernel}} \cup T_{\text{scene}}$ is a structure in $M_{\text{kernel}}$.

**Proof:** We will show that all models of $T_{\text{kernel}} \cup T_{\text{scene}}$ satisfy the properties in the definition of $M_{\text{kernel}}$.

- Property 1 of $M_{\text{kernel}}$ follows easily from axiom 3.34 of $T_{\text{scene}}$.
- Property 2 of $M_{\text{kernel}}$ follows easily from axiom 3.1 of $T_{\text{image}}$. The disjointness of the sets follows from axioms 3.2 - 3.4 of $T_{\text{image}}$.
- Property 3 of $M_{\text{kernel}}$ follows easily from axiom 3.16 of $T_{\text{scene}}$. The disjointness of the sets follows from axioms 3.17 - 3.19 of $T_{\text{scene}}$.

To show that Property 4 of $M_{\text{kernel}}$ is satisfied by all models of $T_{\text{kernel}} \cup T_{\text{scene}}$, we first need to show that any structure which satisfies axioms 3.5 - 3.14 of $T_{\text{image}}$ is an incidence structure without rank. Let $M$ be a model of $T_{\text{kernel}}$. We have already shown that the image elements in $M$ are divided into three disjoint sets. By axioms 3.6 - 3.11, elements are incident iff they are either regions and lines, regions and points, or lines and pixels. Therefore, the image elements in $M$ form an incidence structure. By axiom 3.12 the incidence relation is transitive.

By axioms 3.13, 3.14, and 3.12, every region is incident with a line and a pixel, and hence has type 3. By axiom 3.14, every line is incident with a pixel, and hence has type 2.

To show that Property 5 of $M_{\text{kernel}}$ is satisfied by all models of $T_{\text{kernel}} \cup T_{\text{scene}}$, we need to show that any structure which satisfies axioms 3.20 - 3.33 of $T_{\text{scene}}$ is an incidence structure with rank 3. Let $M$ be a model of $T_{\text{kernel}} \cup T_{\text{scene}}$. We have already shown that the scene elements in $M$ are divided into three disjoint sets. By axioms 3.21 - 3.25, elements are incident iff they are either surfaces and edges, surfaces and points, or edges and points. Therefore, the image elements in $M$ form an incidence structure. By axiom 3.27 the incidence relation is transitive.

By axioms 3.28, 3.29, and 3.32, every surface is incident with an edge and a point, and hence has type 3. By axiom 3.33, every edge is incident with a point, and hence has type 2. By axiom 3.29, every point is incident with an edge and a surface; thus the incidence structure itself has rank 3.

Now consider Property 6 of $M_{\text{kernel}}$. Let $M$ be a model of $T_{\text{kernel}} \cup T_{\text{scene}}$ and let $G$ be a graph whose vertices are the elements in $M$ such that there is an arc between two elements iff for any variable assignment $\sigma$, we have $M, \sigma \models \Delta(i, s)$.
By axiom 3.35, there cannot be an arc between two image elements or between two scene elements, so that the graph $G$ is bipartite.

By axioms 3.36, 3.39, and 3.41, if some node represents a point in $\mathcal{M}$, then there cannot be arcs in $G$ between nodes representing lines or regions.

By axioms 3.37, 3.40 and 3.41, if some node represents an edge in $\mathcal{M}$, then there cannot be arcs in $G$ between nodes representing pixels or regions.

By axioms 3.38, 3.39, and 3.40, if some node represents a surface in $\mathcal{M}$, then there cannot be arcs in $G$ between nodes representing lines or pixels.

Thus $\mathcal{M}$ satisfies condition (6a). It is easy to see that since $\mathcal{M}$ satisfies axiom 3.42, it satisfies condition (6b), and that since $\mathcal{M}$ satisfies axiom 3.43, it satisfies condition (6c).

Property 7 follows from Theorem 3.5. If $\mathcal{M}$ is a model of $T_{\text{kernel}} \cup T_{\text{scene}}$, then it consists of a depicted kernel model $\mathcal{M}_1$ and a set of undepicted surfaces $\mathcal{M}_2$. It is easy to see that $\mathcal{M}_1$ is equivalent to the substructure $C_{\text{dep}}$ and that $\mathcal{M}_2$ is equivalent to $C_{\text{nondep}}$.

Property 8 follows from Theorem 3.7. If $\mathcal{M}$ is a model of $T_{\text{kernel}} \cup T_{\text{scene}}$, then it consists of a depicting kernel model $\mathcal{M}_1$ and a set of nondepicting image objects $\mathcal{M}_2$. It is easy to see that $\mathcal{M}_1$ is equivalent to the substructure $I_{\text{dep}}$ and that $\mathcal{M}_2$ is equivalent to $I_{\text{nondep}}$.

Property 9 follows from Theorem 3.5 and Theorem 3.7.

Property 10 follows from the Depicted Kernel Theorem and the Depicting Kernel Theorem.

Therefore any model of $T_{\text{kernel}} \cup T_{\text{scene}}$ is a structure in $\mathcal{M}_{\text{kernel}}^\text{kernel}$. □

It is important to see the role played in these theorems by axiom 3.16 of $T_{\text{scene}}$. If there were unknown scene objects, we could construct various anomalous models. We would lose Lemma 3.1, and hence the results such as Theorem 3.4 that define how models are constructed piecewise from surfaces, since there are no constraints on unknown scene objects with respect to the composition hierarchy. This would also falsify the Depicted Kernel Theorem, since we could construct models simply by adding undepicted unknown scene objects.

The notion of piecewise construction is also lost if we do not have the coherence axioms 3.28 and 3.29 of $T_{\text{scene}}$, since we could then have undepicted edges without surfaces, or undepicted surfaces without parts that could be added to a model to make an extension.

### 3.4 Assumptions about Images

Although the axiomatization of $T_{\text{kernel}}$ treats image and scene objects and relations in a rather symmetric fashion, in practice we are given an image, and we are trying to construct a model of $T_{\text{kernel}}$ which also satisfies this image. In this case, all models will agree on the image relations as specified by the given image. In particular, for computational reasons, we often make assumptions about the completeness of the theory of the image.

As in [Reiter and Mackworth 89], we will make two important assumptions over the image domain to capture this intuition (see Figure 3.11). First, we assume that we have closure over the image domain, so that the extensions of all image relations is known. Second, we make a unique names assumption over the image domain, so that if $i$ and $i'$ are different constants denoting image objects, then they denote different image objects.
1. **Closed World Assumption for the Image Domain**

All image relations are known.

\[
(\forall x) \ \text{pixel}(x) \equiv x = i_1 \lor \ldots \lor x = i_n
\]

\[
(\forall x) \ \text{line}(x) \equiv x = i_1 \lor \ldots \lor x = i_m
\]

\[
(\forall x) \ \text{region}(x) \equiv x = i_1 \lor \ldots \lor x = i_k
\]

\[
(\forall x, y) \ \text{in}(x, y) \equiv (x = i_1 \land y = i'_1) \lor \ldots \lor (x = i_j \land y = i'_j)
\]

where the \(i_j\) and \(i'_j\) are constants.

2. **Unique Names Assumption for the Image Domain**

All image primitives are pairwise distinct. If \(i\) and \(i'\) are different constants denoting image primitives, they denote different image primitives.

---

**Definition 3.10** *An image is a theory in \(L_{image}\) satisfying \(T_{image} \cup T_{IA}\).*

Since all images must satisfy the closure assumptions of \(T_{IA}\), any image is equivalent to a finite set of ground image literals. \(^4\)

In Figure 3.10(f), the closure axioms for regions, lines, and the \(\text{in}\) relation are:

\[
(\forall x) \ \text{region}(x) \equiv x = r_1 \lor x = r_2 \lor x = r_3
\]

\[
(\forall x) \ \text{line}(x) \equiv x = l_1 \lor \ldots \lor x = l_{14}
\]

\[
(\forall x, y) \ \text{in}(x, y) \equiv (x = l_1 \land y = r_1) \lor (x = l_2 \land y = r_1)
\]

\[
\lor (x = l_3 \land y = r_2) \lor (x = l_4 \land y = r_2) \lor (x = l_5 \land y = r_2)
\]

\[
\lor (x = l_6 \land y = r_3) \lor (x = l_7 \land y = r_3) \lor (x = l_8 \land y = r_3)
\]

\[
\lor (x = l_9 \land y = r_2) \lor (x = l_{10} \land y = r_2) \lor (x = l_{10} \land y = r_3)
\]

\[
\lor (x = l_{11} \land y = r_2) \lor (x = l_{12} \land y = r_2) \lor (x = l_{13} \land y = r_1)
\]

\[
\lor (x = l_{14} \land y = r_2) \lor (x = l_{14} \land y = r_3)
\]

The unique names assumption for regions in this image is

\[
(r_1 \neq r_2) \land (r_1 \neq r_3) \land (r_2 \neq r_3)
\]

---

\(^4\)In the remainder of this work, the symbol \(I\) will denote the set of ground image literals that defines an image.
Chapter 4

An Axiomatic Theory of Shape

Notice that all of the results of the previous chapter are independent of the shape of the surfaces in the scene. In this section, we introduce the relations and axioms that we will need to axiomatize the notion of shape. Several proposals have been made concerning the appropriate representation for shape [Grimson 90], yet none of them have given any independent justification for their particular approaches. In this chapter, the intuitive motivation for shape representation is the question of figure vs. ground — which "side" of an edge is the surface, or figure, and which "side" of the edge is the ground? We will present a set of axioms that formalize these intuitions, and then use these axioms to prove a characterization theorem, namely, that all surfaces have a unique assignment of figure and ground: given the geometric relations among the edges of the surface, there is no ambiguity in this assignment. We will also show these axioms are a minimally sufficient set of conditions to prove this result. This will constitute a theory of shape. In particular, we will present a theory of 2D shapes with straight edges, possibly with holes. In the overall architecture of CardWorld, this chapter investigates the properties of the module $T_{fs}$.

In addition to axiomatizing the class of structures for figure and ground, we also will need to introduce geometric relations, whose axiomatization is beyond the scope of this thesis. Rather than providing explicit axioms for these relations, we will specify their intended interpretations using classes of structures from existing axiomatizations of geometry. This technique will actually be used in each of the remaining CardWorld theories.

4.1 Geometric Relations for Polygonal Surfaces

In this section, we will introduce the set of relations for polygonal surfaces which are concerned with the geometry of the surface. We will not be explicitly axiomatizing these relations within this thesis, since they have been axiomatized elsewhere ([Hilbert 71], [von Plato 95], [von Plato 97], [Tarski 59]). Rather than provide axiomatic proofs for theorems which refer to these geometric relations, we will be using the class of structures that capture the intended interpretations of the relations in the previous axiomatic theories for geometry. In particular, we will be relying upon Hilbert's axiomatization of geometry.

1See Appendix A for a summary of Hilbert's axiomatization of geometry, which we will denote by $T_{Hilbert}$. 

54
Without these relations, it would be possible to have sets of edges in a surface which intuitively do not correspond to a surface. For example, we would not be able to show that the sum of the angles between the edges of a triangle is equal to π, and hence we could have a set of three edges whose relative orientations do not satisfy this property, and hence do not form triangles. On the other hand, the axioms which we present in this chapter for determining figure and ground only require that the extensions of the geometric relations be nonempty. Thus, the axioms of $T_{Fg}$ could still be used to determine figure and ground for sets of edges which violate Euclidean geometry.

The nonlogical lexicon $L_{polygon}$ which we will adopt for polygonal surfaces is:

$$L_{polygon} = L_{scene} \cup \{\theta(a,e_1,e_2), \parallel(e_1,e_2), \text{same}(e_1,e_2,e_3), \text{opposite}(e_1,e_2,e_3), \text{indet}(e_1,e_2), \text{contiguous}(e_1,e_2)\}$$

- The ternary predicate $\theta(a,e_1,e_2)$ denotes the relation whose intended interpretation is that the relative alignment of two non-parallel edges $e_1$ and $e_2$ has the value of $a$ radians.

- The binary predicate $\parallel(e_1,e_2)$ denotes the relation whose intended interpretation is that the edges $e_1,e_2$ are parallel.

- The ternary predicates $\text{same}(e_1,e_2,e_3), \text{opposite}(e_1,e_2,e_3)$ and the binary predicate $\text{indet}(e_1,e_2)$ denote the relations which are defined over the edges in a surface to represent the orientation and relative positions of edges in the surface.

- The binary predicate $\text{contiguous}(e_1,e_2)$ denotes a relation whose intended interpretation is to capture the intuition of convexity — two edges are contiguous iff there are no edges between them.

We will also refer to these predicates as the polygonal relations.

To formally specify the intended interpretations for these predicates, we will define a mapping from the elements in a structure in $M^k_{kernel}$ to elements in the model of Hilbert's axiomatization of geometry:

Definition 4.1 Let $M \in M^k_{kernel}$. Let $h(x)$ be a function which assigns an element in $M$ to an element in the model $H$ of $T_{Hilbert}$, with the following properties:

1. If $(p) \in \text{point}$, then $h(p)$ is a point in $H$.
2. If $(e) \in \text{edge}$, then $h(e)$ is a segment in $H$ that is contained in a polygon.
3. If $(q) \in \text{pixel}$, then $h(q)$ is a point in $H$.
4. If $(l) \in \text{line}$, then $h(l)$ is a segment in $H$.

In the remainder of the thesis, we will interpret the relations in $L_{polygon}$ (and any other required geometric relations) using this mapping and the relations in the model of Hilbert's theory.
4.1.1 Relative Alignment

The ternary relation $\theta(a, e_1, e_2)$ is defined over the edges in a surface to represent the relative alignment of the edges, as in [Grimson 91]. Intuitively, $\theta(a, e_1, e_2)$ iff we can clockwise rotate the line containing the segment for edge $e_1$ by an angle of $a$ radians into the line containing the segment for edge $e_2$, using their intersection point as the pivot. If the intersection point is not part of either segment, then the alignment value is defined modulo $2\pi$ radians. If the intersection point is part of one of the segments, then the alignment value is defined modulo $\pi$ radians. The special case when edges are parallel (i.e. the segments corresponding to the edges do not intersect) is represented by the relation parallel.

In Figure 4.1(a), we have $\theta(a, e_2, e_1)$, where the pivot is the point at which the dashed lines intersect. In Figure 4.1(b) we have indeterminate edges: in this case we have $\theta(a, e_2, e_1)$, where the pivot is the point at which the dashed line intersects the segment for the edge $e_1$.

---

Example: Consider the surface $^2$ represented in Figure 4.2(a). There exists a structure $\mathcal{M}$ in $\mathcal{L}_{\text{polygon}} \cup \mathcal{L}_{\text{scene}}$ with domain $\{e_1, e_2, e_3\}$ for the edges in the surface.

\[
\{\langle \frac{2\pi}{3}, e_1, e_2 \rangle, \langle \frac{\pi}{3}, e_2, e_3 \rangle, \langle \frac{5\pi}{3}, e_3, e_1 \rangle, \langle \frac{\pi}{3}, e_2, e_1 \rangle, \langle \frac{2\pi}{3}, e_1, e_3 \rangle, \langle \frac{\pi}{3}, e_3, e_2 \rangle \} \in \theta
\]

Example: Consider the surface represented in Figure 4.2b. There exists a structure $\mathcal{M}$ in $\mathcal{L}_{\text{polygon}} \cup$

---

$^2$In this chapter, structures in the examples are representing individual surfaces, not scenes with multiple surfaces. It should also be noted that since the theory presented in this chapter is a scene theory, none of the Figures in this chapter are images.
Figure 4.2: Examples of surfaces
\( L_{\text{scene}} \) with domain \( \{e_1, e_2, e_3, e_4, e_5\} \) for the edges in the surface.

\[
\begin{align*}
\{(\frac{\pi}{4}, e_1, e_2) & \quad (\frac{\pi}{4}, e_1, e_3) & \quad (\frac{\pi}{4}, e_5, e_4) & \quad (\frac{\pi}{4}, e_1, e_5) \\
(\frac{\pi}{4}, e_2, e_1) & \quad (\frac{\pi}{4}, e_2, e_3) & \quad (\frac{\pi}{4}, e_2, e_4) & \quad (\frac{\pi}{4}, e_2, e_5) \\
(\frac{\pi}{4}, e_3, e_1) & \quad (\frac{\pi}{4}, e_3, e_2) & \quad (\frac{\pi}{4}, e_3, e_4) & \quad (\frac{\pi}{4}, e_3, e_5) \\
(\frac{\pi}{2}, e_5, e_1) & \quad (\frac{\pi}{2}, e_4, e_2) & \quad (\frac{\pi}{2}, e_4, e_3) & \quad (\frac{\pi}{2}, e_4, e_5) \} \in \theta \\
\{(e_3, e_5) & \quad (e_5, e_3) & \quad (e_1, e_4) & \quad (e_4, e_1) \} \in \text{parallel}
\end{align*}
\]

Note that for the edges \( e_1 \) and \( e_3 \), we have

\[
(\frac{\pi}{2}, e_1, e_3), (\frac{\pi}{2}, e_3, e_1) \in \theta
\]

so that the relative alignment sums to \( \pi \), since the intersection point of the segments corresponding to the edges is contained within the segment for \( e_1 \). □

To handle surfaces which contain edges such as \( e_1 \) and \( e_3 \) in this example, we introduce another binary relation over edges, \( \text{indet}(e_1, e_2) \). Intuitively, one edge \( e_1 \) is indeterminate with respect to another edge \( e_2 \) iff the point at which their corresponding segments in the model of Hilbert’s axiomatization intersect is a point within the segment corresponding to \( e_1 \). An example of a surface with indeterminate edges can be found in Figure 4.2(b), where we have \( (e_1, e_3) \in \text{indet} \).

The intended interpretation of \( \text{indet} \) can be formally specified using Hilbert’s axiomatization of geometry.

**Definition 4.2** Suppose we are given a structure \( \mathcal{M} \in L_{\text{polygon}} \cup L_{\text{scene}} \), and let \( H \) be the model of Hilbert’s axiomatization of geometry.

\[
(\langle e_1, e_2 \rangle \in \text{indet}
\]

iff there exist points \( v_1, v_2 \) such that

\[
(\langle v_1, e_1 \rangle, \langle v_2, e_1 \rangle \in \text{part}
\]

and a point \( p \in H \) such that

\[
(\langle h(v_1), p, h(v_2) \rangle \in \text{ordered}
\]

and

\[
(\langle p, h(e_1) \rangle, \langle p, h(e_2) \rangle \in \text{on}
\]

We will see in the next chapter that indeterminate edges play a special role in the depiction of surfaces.

**Example:** Consider the surface represented in Figure 4.2c, which intuitively is a triangle with a hole.

There exists a structure \( \mathcal{M} \) in \( L_{\text{polygon}} \cup L_{\text{scene}} \) with domain \( \{e_1, e_2, e_3, e_4, e_5, e_6, e_7, e_8\} \) for the edges in the surface.

\[
(\langle \frac{\pi}{2}, e_1, e_2 \rangle, \langle \frac{\pi}{2}, e_2, e_3 \rangle, \langle \frac{\pi}{2}, e_3, e_1 \rangle, \langle \frac{\pi}{2}, e_4, e_5 \rangle) \in \theta
\]
More formally, we can define $\theta$ and parallel with respect to Hilbert’s axiomatization of geometry:

**Definition 4.3** Suppose we are given a structure $M \in L_{\text{polygon}} \cup L_{\text{scene}}$, and let $\mathcal{H}$ be the model of Hilbert’s axiomatization of geometry.

We have the following:

- $\langle a, e_1, e_2 \rangle \in \theta$ iff
  
  $\langle \tilde{h}(e_1), \tilde{h}(e_2) \rangle \in \text{angle}$

  $a = \text{magnitude}(\tilde{h}(e_1), \tilde{h}(e_2))$

- If $\langle e_1, e_2 \rangle \notin \text{indet}$, then $0 < a < 2\pi$. If $\langle e_1, e_2 \rangle \in \text{indet}$, then $0 < a < \pi$.

- $\langle e_1, e_2 \rangle \in \text{parallel}^M$ iff $\langle \tilde{h}(e_1), \tilde{h}(e_2) \rangle \in \text{parallel}^H$, i.e., edges are parallel in $M$ iff the segments in $\mathcal{H}$ corresponding to the edges are also parallel in $\mathcal{H}$.

Note that $\theta$ is only defined for edges whose segments in $\mathcal{H}$ intersect. Also, since the value of the alignment must be greater than 0, “overlapping” edges are not allowed.

### 4.1.2 Positional Relations

The following example shows that the alignment relation alone will be insufficient for characterizing figure and ground, since we have two distinct surfaces which are equivalent with respect to the extension of $\theta$, yet which intuitively have different shapes.
Example: Consider the surfaces represented in Figure 4.3. In both cases there is a structure $\mathcal{M}$ in $\mathcal{L}_{\text{polygon}} \cup \mathcal{L}_{\text{scene}}$ such that \{e$_1$, e$_2$, e$_3$\} $\subset M$ and

\[
\langle \frac{\pi}{4}, e_1, e_2 \rangle, \langle \frac{3\pi}{2}, e_1, e_3 \rangle, \langle \frac{\pi}{4}, e_2, e_3 \rangle, \langle \frac{7\pi}{4}, e_2, e_1 \rangle, \langle \frac{\pi}{2}, e_3, e_1 \rangle, \langle \frac{7\pi}{4}, e_3, e_2 \rangle \in \theta
\]

In order to be able to distinguish between these two cases, we introduce the ternary relations \textit{sameside} and \textit{opposite}. Implicit in the notion of relative alignment is that the line corresponding to an edge "divides" the plane into two sides, and we want to determine whether two other edges are on the same side or opposite sides of the plane. If the two edges are on the same side of the plane, then one edge will be in the subset of the plane bounded by the other two edges. If the two edges are not on the same side, then one edge will not be in a subset of the plane bounded by the other two edges.

We can formally specify the intended interpretations of the positional relations using models of Hilbert's geometry. We will first need to introduce the notion of an edge being bounded by two other edges. There are four cases here, depending on whether the relative alignments of the edges. If the measures of the angles $\angle(e_3, e_1)$ and $\angle(e_3, e_2)$ are both less than $\pi$, we will say that an edge $e_1$ is bounded by two other edges $e_2$ and $e_3$ iff the interior of the angle $\angle(e_3, e_1)$ has a nonempty intersection with the interior of the angle $\angle(e_3, e_2)$. If the measure of the angle $\angle(e_3, e_1)$ is less than $\pi$ and the measure of the $\angle(e_3, e_2)$ is greater than $\pi$, we will say that an edge $e_1$ is bounded by two other edges $e_2$ and $e_3$ iff the interior of the angle $\angle(e_3, e_1)$ has a nonempty intersection with the exterior of the angle $\angle(e_3, e_2)$. If the measures of the angles are equal to $\pi$ or if the edges are parallel, then $e_1$ is bounded by $e_2$ and $e_3$ iff it is between $e_2$ and $e_3$. (A formal definition for this relation can be found in Appendix A).

\textbf{Definition 4.4} Suppose we are given a structure $\mathcal{M} \in \mathcal{L}_{\text{polygon}} \cup \mathcal{L}_{\text{scene}}$, and let $\mathcal{H}$ be the model of Hilbert's axiomatization of geometry. 

\[
\langle e_1, e_2, e_3 \rangle \in \text{sameside}
\]

iff either $e_1$ is bounded by $e_2$ and $e_3$ or $e_2$ is bounded by $e_1$ and $e_3$.

\[
\langle e_1, e_2, e_3 \rangle \in \text{opposite}
\]

iff $e_1$ is not bounded by $e_2$ and $e_3$, and $e_2$ is not bounded by $e_1$ and $e_3$.

Example: Consider the surface represented in Figure 4.2a. There exists a structure $\mathcal{M}$ in $\mathcal{L}_{\text{polygon}} \cup \mathcal{L}_{\text{scene}}$ with domain \{e$_1$, e$_2$, e$_3$\} $\subset M$ for the edges in the surface such that

\[
\langle e_2, e_3, e_1 \rangle, \langle e_1, e_3, e_2 \rangle, \langle e_1, e_2, e_3 \rangle \in \text{opposite}
\]
Example: Consider the surface represented in Figure 4.2b. There exists a structure \( M \) in \( L_{\text{polygon}} \cup L_{\text{scene}} \) with domain \( \{e_1, e_2, e_3, e_4, e_5\} \subseteq M \)

\[(e_2, e_3, e_1), (e_2, e_4, e_1), (e_2, e_5, e_1), (e_1, e_3, e_2), (e_1, e_5, e_2), (e_3, e_4, e_2), (e_4, e_5, e_2) \in \text{opposite} \]

\[(e_3, e_4, e_1), (e_3, e_5, e_1), (e_4, e_5, e_1), (e_1, e_4, e_2), (e_3, e_5, e_2) \in \text{sameside} \]

\[\square\]

Example: Consider the surface represented in Figure 4.2c. There exists a structure \( M \) in \( L_{\text{polygon}} \cup L_{\text{scene}} \) with domain \( \{e_1, e_2, e_3, e_4, e_5, e_6, e_7, e_8 \in M\} \) for the edges in the surface such that

\[(e_2, e_3, e_1), (e_1, e_3, e_2), (e_1, e_2, e_3), (e_5, e_6, e_4), (e_4, e_6, e_5), (e_4, e_8, e_5), (e_5, e_7, e_4), (e_5, e_8, e_4) \in \text{opposite} \]

\[(e_6, e_7, e_4), (e_8, e_8, e_4), (e_7, e_8, e_4), (e_4, e_7, e_5), (e_4, e_7, e_2), (e_4, e_8, e_2), (e_5, e_6, e_2), (e_5, e_7, e_2) \in \text{sameside} \]

\[\square\]

### 4.1.3 Contiguous Edges

The intuition for contiguous edges is based on the intuition of “betweenness” in Hilbert’s geometry – two edges are contiguous if we can draw a segment between the two edges which does not cross any other edges in the surface. For example, for the surface in Figure 4.2(b), the following pairs of edges are contiguous –

\[(e_5, e_4), (e_5, e_1), (e_3, e_4), (e_5, e_2) \in \text{contiguous} \]

but \( (e_5, e_3) \not\in \text{contiguous} \), since any segment between \( e_5 \) and \( e_3 \) will intersect with \( e_4 \).

We can formally characterize the intended interpretation of contiguous as follows:

**Definition 4.5** Suppose we are given a structure \( M \in L_{\text{polygon}} \). Let \( H \) be the model of Hilbert's geometry.

\[\langle e_1, e_2 \rangle \in \text{contiguous}\]

iff there exist points \( p_1, p_2 \in M \) such that

\[\langle p_1, e_1 \rangle \in \text{part}, \langle p_2, e_2 \rangle \in \text{part}\]

and there does not exist an element \( x \) such that

\[\langle h(p_1), x, h(p_2) \rangle \in \text{ordered} \]
4.1.4 Structures for $L_{\text{polygon}}$

Having specified the intended interpretations of the predicate symbols in $L_{\text{polygon}}$, we can now define a class of structures which pulls all of this together.

**Definition 4.6** $M_{i}^{\text{polygon}}$ is the class of structures in $L_{\text{polygon}} \cup L_{\text{scene}}$ in which the extensions of $\theta$ and parallel are specified in Definition 4.3, the extensions of same side and opposite are specified in Definition 4.4, the extension of indet is specified in Definition 4.2, and the extension of contiguous is specified by Definition 4.5.

We can also characterize the extensions of the relations in $L_{\text{polygon}}$ more explicitly. This will be essential later in the chapter when we use these structures to specify the intended models for figure/ground assignments.

Given a surface, we will be interested in structures which are isomorphic to the extension of contiguous for different sets of edges. In particular, we will be interested in sets of edges whose relative alignment is an angle whose value is less than $\pi$.

**Definition 4.7** Given a structure $M$ in $L_{\text{polygon}}$, the contiguous graph is a directed graph $G = (N, A)$ with the following properties:

1. The set $N$ of nodes of the graph is the set of edges in the surface.
2. $(e_1, e_2)$ is an arc in $A$ iff $$(e_1, e_2) \in \text{contiguous}$$ and either $(e_1, e_2) \in \text{parallel}$, or $(\alpha, e_1, e_2) \in \theta$ and $\alpha < \pi$.

The contiguous graph for the surface in Figure 4.2(a) can be found in Figure 4.4(a). In this case, all edges are contiguous with each other. The contiguous graph for the surface in Figure 4.2(b) can be found in Figure 4.4(b). There is no arc between $e_5$ and $e_3$ because they are not contiguous; there is no directed arc between $e_4$ and $e_3$ because the relative alignment is greater than $\pi$.

In reasoning about the extensions of opposite and same side, the following sentences will be useful in later sections, where they will be used to prove the completeness of various model constructions. They are satisfied in all structures in $M_{i}^{\text{polygon}}$:

$$(\forall e_1, e_2, e_3, e_4) \ sameside(e_1, e_2, e_3) \land \ sameside(e_4, e_1, e_3) \supset \ sameside(e_2, e_4, e_3)$$
\[(\forall e_1, e_2, e_3) \text{sameside}(e_1, e_2, e_3) \equiv \text{sameside}(e_2, e_1, e_3)\]
\[(\forall e_1, e_2, e_3, e_4) \text{opposite}(e_1, e_2, e_3) \land \text{opposite}(e_1, e_4, e_3) \supset \text{sameside}(e_2, e_4, e_3)\]
\[(\forall e_1, e_2, e_3, e_4) \text{sameside}(e_1, e_2, e_3) \land \text{opposite}(e_2, e_4, e_3) \supset \text{opposite}(e_1, e_4, e_3)\]
\[(\forall e_1, e_2, e_3) \text{opposite}(e_1, e_2, e_3) \equiv \text{opposite}(e_2, e_1, e_3)\]
\[(\forall e_1, e_2) \text{indet}(e_1, e_2) \supset \neg \text{indet}(e_2, e_1)\]

Using these sentences, we can show that the extensions of \text{sameside} and \text{opposite} within a structure of \(\mathcal{M}_i^{\text{polygon}}\) is isomorphic to the following class of graphs:

**Definition 4.8** \(K_{1,n}\) denotes the complete bipartite graph \(G = (X, Y, E)\) with \(|X| = 1, |Y| = n,\) and all pairs between \(X\) and \(Y\) being edges.

**Definition 4.9** A butterfly graph \(G = (V_1 \cup V_2 \cup \{x\}, A)\) is a \(K_{1,n}\) graph with root node \(x\) and whose remaining nodes are partitioned into two disjoint sets \(V_1\) and \(V_2\).

**Theorem 4.1** In any structure \(\mathcal{M} \in \mathcal{M}_i^{\text{polygon}}\), the extension of \text{sameside} and \text{opposite} for a surface is isomorphic to the union of a set of butterfly graphs for each edge in the surface, such that for each edge \(e_1\) in the surface

1. \(e_1\) is the root node of a butterfly graph;
2. \((e_1, e_2) \in A\) iff \(e_2 \in V_1\) or \(e_2 \in V_2\);
3. \((e_2, e_3, e_1) \in \text{sameside}\) iff \(e_2, e_3 \in V_1\) or \(e_2, e_3 \in V_2\);
4. \((e_2, e_3, e_1) \in \text{opposite}\) iff \(e_2 \in V_1\) and \(e_3 \in V_2\).

**Proof:** We will use the sentences above which are entailed by structures in \(\mathcal{M}_i^{\text{polygon}}\).

Let \(\mathcal{M} \in \mathcal{M}_i^{\text{polygon}}\). Construct a graph \(G = (N, A_1)\) such that \((e_2, e_1) \in A_1\) iff
\[(e_2, e_3, e_1) \in \text{sameside}\]

It is easy to see that this graph will be a \(K_{1,n}\) with root node \(e_1\).

We next need to show that the nodes of \(G\) can be partitioned into two disjoint sets. For any edge \(e\), let
\[V_1(e) = \{e' : (e, e', e_1) \in \text{sameside}\}\]
\[V_2(e) = \{e' : (e, e', e_1) \in \text{opposite}\}\]

Since \text{sameside} is symmetric and transitive in its first two arguments, we have for any edges \(e_2, e_3\), if \((e_1, e_2, e_3) \in \text{sameside}\), then \(V_1(e_2) = V_1(e_3)\). Since the extensions of \text{sameside} and \text{opposite} are disjoint, we have \(V_1(e_2) \cap V_2(e_2) = \emptyset\).

Thus, the graph \(G\) is a butterfly graph. \(\square\)

The butterfly graphs for the edges in the surface in Figure 4.2(a) can be found in Figure 4.5. The butterfly graphs for the edges in the surface in Figure 4.2(b) can be found in Figure 4.6.

Note that not every butterfly graph is isomorphic to the extension of \text{sameside} and \text{opposite} within a structure in \(\mathcal{M}_i^{\text{polygon}}\). This is equivalent saying that not every butterfly graph corresponds to the extension of \text{sameside} and \text{opposite} in a physically realizable surface. However, the definition of \(\mathcal{M}_i^{\text{polygon}}\) uses the model of Hilbert's axiomatization of geometry to guarantee that the extensions of these relations do indeed satisfy their intended interpretations, so that we may assume that all positional graphs which are substructures of our models will correspond to physically intuitive surfaces.
Figure 4.5: Butterfly graphs isomorphic to the extension of *sameside* and *opposite* for the surface in Figure 4.2(a).

Figure 4.6: Butterfly graphs isomorphic to the extension of *sameside* and *opposite* for the surface in Figure 4.2(b).
4.2 Figure and Ground: Intuitions

To this point, we have specified a set of scene relations for the edges in a surface. As of yet, these axioms are independent of any notion of figure and ground, which we have proposed is essential in any theory of shape. We now present a set of axioms that formalize the relationship between these geometric and scene relations and the intuitions we have about figure and ground.

In this section, we will introduce the intuitions behind the figure-ground relations and give different examples of classes of surfaces together with their intuitive figure-ground assignments. In the next section, we will formally characterize the class of structures which are the intended interpretations of the figure-ground relations, and provide a correct and complete axiomatization for this class of structures in the theory $T_{fg}$.

4.2.1 Language of $T_{fg}$

The language which we will adopt for surfaces is a second-order language with equality, with the following nonlogical lexicon:

$$L_{fg} = L_{kernel} \cup L_{polygon}$$

$$\cup \{meet(e_1, e_2, v), connected(e_1, e_2), outer(e),$$

$$\quad inside(e_1, e_2), outside(e_1, e_2)\}$$

• The ternary predicate $meet(e_1, e_2, v)$ denotes the relation defined over two edges and a point, and whose intended interpretation is that the two edges meet at a vertex point within the surface.

• The binary predicate $connected(e_1, e_2)$ is defined over the set of edges in a surface. The intended interpretation is that two edges are connected iff there exists a linear sequence of edges from one to the other such that every two consecutive edges in this sequence meet.

• The unary predicate $outer(e)$ defines a set of edges. Intuitively, the edges in a surface can be partitioned into edges which form a hole in the surface and edges which form the boundary of the surface. The intended interpretation for this relation is that edges which form a hole are not outer, while edges which form the boundary of the surface are outer edges.

• The binary predicate $inside(e_1, e_2)$ is defined over the edges in a surface. The intended interpretation is that $inside(e_1, e_2)$ iff the surface is on the "$e_2$ side" of $e_1$ in a clockwise direction.

• The binary predicate $outside(e_1, e_2)$ is defined over the edges in a surface. The intended interpretation is that $outside(e_1, e_2)$ iff the surface is on the "$e_2$ side" of $e_1$ in a counterclockwise direction.

• Since the language is second-order, there is also a countable set of predicate variables of all arities.
4.2.2 Motivating Examples: What is a Surface?

Within $T_{scene}$, we introduced the scene objects – surfaces, edges, and points. The axioms of defined surfaces as sets of edges and points which satisfied the part relation. In this section, we extend the characterization of surfaces by considering three additional intuitions.

Our intuitions about surfaces deal with how edges within the same surface are related to each other.

**Intuition 1** The set of edges in a surface are not an arbitrary set of edges; rather, each edge in a surface is incident with exactly two other distinct edges in the surface.

We will say that two edges meet at a vertex iff they have a unique point (i.e. the vertex) which is a part of both edges. Thus, edges that meet are part of the same surface.

Consider the surface in Figure 4.2(b). There exists a structure $M$ in $L_{fg}$ with domain 
$$ \{e_1, e_2, e_3, e_4, v_1, v_2, v_3, v_4, v_5\} \subset M $$
such that
$$ \langle e_1, e_2, v_1 \rangle, \langle e_2, e_3, v_2 \rangle, \langle e_3, e_4, v_3 \rangle, \langle e_4, e_5, v_4 \rangle, \langle e_5, e_1, v_5 \rangle \in \text{meet} $$

If we consider the sets of edges in Figure 4.7, none of these should be surfaces. In (a) and (b), there exist edges which do not meet other edges. In (b) and (c), there are edges which meet more than two other edges.

**Intuition 2** All edges in a surface are organized into sets of connected cycles.

Consider the surface in Figure 4.2(b). There exists a structure $M$ in $L_{fg}$ with domain $\{e_1, e_2, e_3, e_4, e_5\} \subset M$ such that the surface in has one set of connected edges consisting of $\{e_1, e_2, e_3, e_4, e_5\}$. If we consider
the surface in Figure 4.2(c), there exists a structure \( M \) in \( \mathcal{L}_{fg} \) with domain \( \{e_1, e_2, e_3, e_4, e_5, e_6, e_7, e_8\} \subseteq M \) such that the surface has two disjoint sets of connected edges:

\[
\{e_1, e_2, e_3\}, \{e_4, e_5, e_6, e_7, e_8\}
\]

The relation \( \text{connected}(e_1, e_2) \) specifies that two edges are members of the same cycle of edges within the surface.

**Intuition 3** Surfaces may contain holes, in which case there are multiple disjoint cycles of connected edges.

The relation \( \text{outer} \) distinguishes between those edges that form the boundaries of holes in the surface, and those edges which do not. All outer edges are in the same connected cycle: there cannot be two different connected cycles of outer edges. Therefore each surface contains a unique set of outer edges.

For example, consider the surface in Figure 4.2b. There exists a structure \( M \) in \( \mathcal{L}_{fg} \) such that \( \{e_1, e_2, e_3, e_4, e_5\} \subseteq M \) and

\[
\langle e_1 \rangle, \langle e_2 \rangle, \langle e_3 \rangle, \langle e_4 \rangle, \langle e_5 \rangle \in \text{outer}
\]

since there are no holes in the surface.

Consider the surface in Figure 4.2c. There exists a structure \( M \) in \( \mathcal{L}_{fg} \) such that \( \{e_1, e_2, e_3, e_4, e_5, e_6, e_7, e_8\} \subseteq M \) and

\[
\langle e_1 \rangle, \langle e_2 \rangle, \langle e_3 \rangle \in \text{outer}
\]

\[
\langle e_4 \rangle, \langle e_5 \rangle, \langle e_6 \rangle, \langle e_7 \rangle, \langle e_8 \rangle \notin \text{outer}
\]

since the edges \( e_4, e_5, e_6, e_7, e_8 \) form the boundary of a hole in the surface.

We can summarize the distinction between outer edges and non-outer edges with the following definitions:

**Definition 4.10** Given a structure \( M \) in \( \mathcal{L}_{fg} \), an outer edge is an edge in the following set:

\[
\{e \mid e \in M, \langle e \rangle \in \text{outer}\}
\]

**Definition 4.11** Given a structure \( M \) in \( \mathcal{L}_{fg} \), a hole edge is an edge in the following set:

\[
\{e \mid e \in M, \langle e \rangle \notin \text{outer}\}
\]

In the example for the surface in Figure 4.2(c), the edges \( \{e_1, e_2, e_3\} \) are the outer edges and the edges \( \{e_4, e_5, e_6, e_7, e_8\} \) are the hole edges.

### 4.2.3 Motivating Examples: Figure-Ground Assignments

The problem of figure and ground is to decide which side of an edge is the surface and which side of the edge is not the surface.
Figure 4.8: Figure ground assignments for the example surfaces.
Intuition 4 *Surfaces lie on a unique "side" of an edge.*

For example, if we consider the surfaces in Figure 4.2, the intuitive figure ground assignments are represented in Figure 4.8, where the arrows indicate on which side of the edge lies the surface.

Every edge has two sides, so we will define the “figure” side of an edge using the binary relations *inside* and *outside*. Intuitively, these relations are related to the relative alignment between edges – *inside*(e₁, e₂) iff the surface is on the “e₂ side” of e₁ in a clockwise direction (using the intersection point of their corresponding segments as the pivot) and *outside*(e₁, e₂) iff the surface is on the “e₂ side” of e₁ in a counterclockwise direction. Thus *inside* is similar to the clockwise convexity of [Huttenlocher and Wayner 92] and *outside* is similar to counterclockwise convexity.

The problem with which we are faced is to determine the extensions of *inside* and *outside*, given a surface with a complete assignment of alignment and positional relations.

**Definition 4.12** A *figure ground assignment* in a structure M in L₉ is the substructure consisting of the extensions of the relations *inside* and *outside* in M.

Given any two edges in the surface, we must be able to determine their figure/ground assignment. To do this, we will consider arbitrary edges which satisfy the different polygon relations in a surface, and use these to illustrate the intended intuitive figure/ground assignment for these edges. We will consider the intuitive figure/ground assignments for each of the following cases:

- two edges which meet
- two edges which are on the same side of another edge
- two edges which are on opposite sides of another edge
- indeterminate edges
- hole edges
- anomalous figure/ground assignments

In the next section, we will formalize each class of these intuitive figure/ground assignments as classes of structures for the extensions of the *inside* and *outside* relations.

**Edges which meet:** Consider two arbitrary edges which meet at some vertex (see Figure 4.9). Intuitively, in some structure M in L₉, there are only two possible figure-ground assignments:

\[ (e₁, e₂) \in \text{inside}, \ (e₁, e₂) \notin \text{outside} \]

\[ (e₂, e₁) \notin \text{inside}, \ (e₂, e₁) \in \text{outside} \]

or

\[ (e₁, e₂) \notin \text{inside}, \ (e₁, e₂) \in \text{outside} \]
Figure 4.9: Figure/ground assignments for edges which meet.

\[(e_2, e_1) \in \text{inside}, (e_2, e_1) \notin \text{outside}\]

In the first case, the edges form a convex corner in the surface, while in the second case, they form a concave corner in the surface.

**Edges which are sameside:** In Figure 4.10 we have three arbitrary edges which satisfy

\[(e_1, e_2, e_3) \in \text{sameside}\]

for some structure \(\mathcal{M}\) in \(L_{f_3}\). Intuitively, there are only two possible figure-ground assignments for \(e_3\) and the two edges \(e_1, e_2\):

\[(e_3, e_1) \in \text{inside}, (e_3, e_1) \notin \text{outside}\]

\[(e_3, e_2) \in \text{inside}, (e_3, e_2) \notin \text{outside}\]

or

\[(e_3, e_1) \notin \text{inside}, (e_3, e_1) \in \text{outside}\]

\[(e_3, e_2) \notin \text{inside}, (e_3, e_2) \in \text{outside}\]

**Edges which are opposite:** In Figure 4.11 we have three arbitrary edges which satisfy

\[(e_1, e_2, e_3) \in \text{opposite}\]

for some structure \(\mathcal{M}\) in \(L_{f_3}\). Intuitively, there are only two possible figure-ground assignments for \(e_3\) and the two edges \(e_1, e_2\):

\[(e_3, e_1) \in \text{inside}, (e_3, e_1) \notin \text{outside}\]

\[(e_3, e_2) \notin \text{inside}, (e_3, e_2) \in \text{outside}\]
Figure 4.10: Figure/ground assignments for edges which are on the same side as another edge.

Figure 4.11: Figure/ground assignments for edges which are on opposite sides of another edge.

or

\[ (e_3, e_1) \notin \text{inside}, (e_3, e_1) \in \text{outside} \]

\[ (e_3, e_2) \in \text{inside}, (e_3, e_2) \notin \text{outside} \]

**Indeterminate Edges** Although the intended interpretation of the relation *inside* is similar to the clockwise convexity of [Huttenlocher and Wayner 92] and *outside* is similar to counterclockwise convexity, there are sets of edges which Huttenlocher and Wayner's definition do not cover, namely the indeterminate, or *indet*, edges. For these edges, the surface is both inside an edge and outside the edge. In Figure 4.12, we illustrate the two cases of *indet* edges; the arrows indicate on which side of the edges is the surface. In all of the surfaces in this figure, we have \( (e_2, e_1) \in \text{indet} \).
Figure 4.12: Figure/ground assignments among indeterminate edges.
Notice that for non-indeterminate edges (i.e. \( \langle e_2, e_1 \rangle \not\in \text{indet} \)), in any structure \( \mathcal{M} \) we have either

\[
\langle e_2, e_1 \rangle \in \text{inside}, \langle e_2, e_1 \rangle \not\in \text{outside}
\]

or

\[
\langle e_2, e_1 \rangle \not\in \text{inside}, \langle e_2, e_1 \rangle \in \text{outside}
\]

while for indeterminate edges (i.e. \( \langle e_2, e_1 \rangle \in \text{indet} \)), we have either

\[
\langle e_2, e_1 \rangle \in \text{inside}, \langle e_2, e_1 \rangle \in \text{outside}
\]

as in Figure 4.12(a) and Figure 4.12(c), or

\[
\langle e_2, e_1 \rangle \in \text{inside}, \langle e_2, e_1 \rangle \in \text{outside}
\]

as in Figure 4.12(b) and Figure 4.12(d). Thus, we will want to include the condition that for indeterminate edges, the extensions of \text{inside} and \text{outside} are isomorphic, while for edges which are not indeterminate, the extension of \text{inside} is isomorphic to the inverse of the extension of \text{outside}.

We can combine all of these intuitions to determine the figure-ground assignments for some entire surfaces.

**Example:** Consider the surface in Figure 4.2a. There exists a structure \( \mathcal{M} \) in \( \mathcal{L}_{fg} \) with domain \( \{e_1, e_2, e_3\} \subseteq \mathcal{M} \) such that

\[
\langle e_1, e_2 \rangle \not\in \text{inside}, \langle e_1, e_2 \rangle \in \text{outside}; \langle e_2, e_3 \rangle \not\in \text{inside}, \\
\langle e_2, e_3 \rangle \in \text{outside}; \langle e_3, e_1 \rangle \not\in \text{inside}, \langle e_3, e_1 \rangle \in \text{outside}, \\
\langle e_2, e_1 \rangle \in \text{inside}, \langle e_2, e_1 \rangle \not\in \text{outside}; \langle e_3, e_2 \rangle \in \text{inside}, \\
\langle e_3, e_2 \rangle \not\in \text{outside}; \langle e_1, e_3 \rangle \in \text{inside}; \langle e_1, e_3 \rangle \not\in \text{outside}
\]
We can represent the extensions of inside and outside in a structure $\mathcal{M}$ as directed graphs in which the edges of the surface are the nodes of the graph, and there is a directed arc $(e_1, e_2)$ in the graph iff $(e_1, e_2) \in \text{inside}$ (the inside graph), or $(e_1, e_2) \in \text{outside}$ (the outside graph). The inside and outside graphs for the preceding example are found in Figure 4.13. Note that there are no indeterminate edges in this surface, so that the inside and outside graphs for the surface are inverses of each other.

**Example:** Consider the surface in Figure 4.2(b). There exists a structure $\mathcal{M}$ in $\mathcal{L}_{f_2}$ with domain 
\{e_1, e_2, e_3, e_4, e_5\} $\subseteq M$ such that the extension of inside is isomorphic to the graph in Figure 4.14(a), and the extension of outside is isomorphic to the graph in Figure 4.14(b).

Note that this model satisfies 
\[(e_1, e_3), (e_2, e_4) \in \text{indet}\]

so that these edges are preserved in both the inside and outside graphs. □

**Holes and Surfaces** Consider again the surface in Figure 4.2(c). It is interesting to note that the surface in the model contains a connected set of edges which is isomorphic to the one in Figure 4.2b, yet which has the inverse figure/ground assignment i.e., the inside and outside graphs for this set of edges
is the inverse of the inside graph for the set of edges in Figure 4.2(b) (see Figure 4.15). The set of edges in this surface form a hole, and this leads to the following intuition:

**Intuition 5** Hole edges have inverse figure/ground assignments from sets of outer edges.

In fact, for any connected set of edges, there should exist a surface in which these are outer edges, and there should also exist a surface in which these are hole edges. (Note that this is the case in Figure 4.2(a) and 4.2(b)).

**Eliminating Unintuitive Figure/Ground Assignments** In addition to capturing the intuitive figure/ground assignments for surfaces, we also want to rule out nonintuitive figure/ground assignments for surfaces.

In particular, we have the following:

**Intuition 6** Surfaces have unique figure/ground assignments.

Thus, not all combinations of inside graphs for the different closed contours of edges in a surface will correspond to an intuitive figure/ground assignment. If a surface has \( n \) different closed contours of edges, then the surface could have \( 2^n \) possible figure/ground assignments; yet intuitively, every surface has a unique figure/ground assignment.

However, even if we can provide a theory of figure and ground which gives a surface a unique figure/ground assignment, we still need to justify whether it really is an intuitively correct assignment. For example, in the figure/ground assignment for the surface in Figure 4.16(a), the edges form the boundary of a hole, yet there is no boundary to the surface. Such a figure/ground assignment cannot possibly correspond to a surface.

The figure/ground assignments in Figure 4.17 (b), (c), and (d) are the three unintuitive assignments for a surface with one hole. The set of edges in Figure 4.17 (e) intuitively cannot be a surface, and hence should not have any consistent figure/ground assignment. The problem in this case is that there seem to be too many nested cycles of edges, so that we cannot determine which should be the outer edges and which edges belong to a hole in the surface.

Why are the examples in Figure 4.17 and Figure 4.16 unintuitive?

**Intuition 7** The edges of a surface form two sides of the surface, and the surface is between these edges, so that every edge has a corresponding outer edge which intuitively is the "other side" of the surface.

We can think of the two sides of a surface in terms of the figure/ground assignments. It must at least be the case that for these two edges we have \( (e_1, e_2) \in \text{inside} \) and \( (e_2, e_1) \in \text{outside} \). That is, the figure is on the clockwise side of one edge and the counterclockwise side of the other edge.

Again, consider how the two unintuitive figure ground assignment for the surfaces in Figure 4.16 violate this intuition. In Figure 4.16 (a), even though all of the edges are outer edges, there are no edges which are on the "other" side of the surface from either \( e_1, e_2, \) or \( e_3 \). In (b), \( e_4 \) should intuitively be
Figure 4.16: Anomalous figure ground assignments
Figure 4.17: Additional unintuitive figure ground assignments.
the "other side" of the surface for the edges $e_1$ and $e_2$, yet in the unintuitive assignment, the surface is on the wrong side of the edge $e_4$.

How can we make the intuition about the "other side" of the surface more precise? First of all, the two edges must be contiguous, so that there are no other edges between them. For example, edge $e_5$ could not be the "other side" of the surface for edge $e_1$ because there are other edges between them. Secondly, the two edges are intuitively convex: that is, their figure/ground assignment is based on their relative alignment, so that the "figure" side of the edge is the one in which the alignment is less than $\pi$.

In Figure 4.16(a), the only edges which are both contiguous and convex with respect to the edge $e_2$ are the edges $e_1$ and $e_3$, and hence they should be the edges which are on the "other side" of the surface. In Figure 4.16(b), the edge $e_4$ is both contiguous and convex with respect to the edge $e_1$ and hence should be the other side: however, the figure/ground assignment disagrees with this, and hence is anomalous.

4.3 Axiomatization of Figure and Ground

In this section, we formalize the intuitions about figure/ground assignments which we have just considered. We will first define various classes of structures related to the extension of each relation in $L_{fg}$, before specifying the class of structures for the intended interpretations of $T_{fg}$. We then present the axioms of $T_{fg}$ and show that the structures that we have defined do indeed satisfy the axioms of $T_{fg}$. Finally, we provide an axiomatizability theorem for the models of $T_{fg}$.

4.3.1 Structures for Figure and Ground

Given the intuitions and examples of the previous section, we can now formally characterize the structures in $M_{fg}$ which we will later show to be the set of models for $T_{scene} \cup T_{fg}$. The challenge of this section is to formally characterize the intuitive figure/ground assignments using different classes of structures. Each of the classes of structures which we define will either be purely combinatorial (such as classes of graphs) or they will be structures which are satisfied by the intended interpretations of the relations in $L_{polygon}$ (e.g. contiguous graphs and butterfly graphs). The objective is to specify these classes so that any structure isomorphic to one in the class will be isomorphic to a model of the axioms of $T_{fg}$ and also will correspond to an intuitive figure/ground assignment or other properties of surfaces. A summary of the classes of structures defined in this section can be found in Table 4.1.

The first class of structures we will consider will be those which capture the intended interpretation of the relations meet and connected. Since we will be characterizing the extensions of these relations using different classes of graphs, we will need a few introductory definitions:

**Definition 4.13** A $C_n$ graph is a connected set of $n$ nodes in which each edge is adjacent to exactly two other nodes.

**Definition 4.14** A directed $C_n$ is a directed subgraph of a $C_n$ graph with a unique cycle of length $n$. 

<table>
<thead>
<tr>
<th>Class of structures</th>
<th>Relation</th>
</tr>
</thead>
<tbody>
<tr>
<td>contour graphs</td>
<td>extension of <em>meet</em> and <em>connected</em></td>
</tr>
<tr>
<td>connected figure/ground graphs</td>
<td>extension of <em>inside</em> for connected edges</td>
</tr>
<tr>
<td>contiguous figure/ground graphs</td>
<td>extension of <em>inside</em> for hole edges</td>
</tr>
<tr>
<td>surface figure/ground graphs</td>
<td>extension of <em>inside</em> for a surface</td>
</tr>
<tr>
<td>inverse figure/ground graphs</td>
<td>extension of <em>outside</em> for a surface</td>
</tr>
</tbody>
</table>

Table 4.1: Classes of structures for characterizing figure/ground assignments.

With these definitions in hand, we can characterize the extension of *meet* as a class of $C_n$ graphs:

**Definition 4.15** Given a structure $M$ in $L_{fg}$, a contour graph $G = (N, A)$ is a $C_n$ graph such that $N$ is a set of edges in $M$, $n = |N|$, and $(e_1, e_2) \in A$ iff there exists a point $p \in M$ such that

$$\langle p, e_1, \langle p, e_2 \rangle \in \text{part}$$

**Example:** The contour graph for the surface in Figure 4.2(b) is shown in Figure 4.18(a). The contour graph for the surface in Figure 4.2(c) is shown in Figure 4.18(b). Notice that the contour graph in (b) has two disconnected subgraphs, which correspond to the two different sets of connected edges in the surface. □

The second class of structures will be used to characterize the possible figure/ground assignments for connected sets of edges. This will be done by defining subgraphs related to the different polygon relations, and then combining these subgraphs to create a structure which is isomorphic to the extension of *inside* for a connected set of edges. We have already defined the class of contour graphs related to the extensions of *meet* and *connected*. We next define the class of subgraphs related to the extensions of the positional relations.

**Definition 4.16** Given a butterfly graph $G = (V_1 \cup V_2 \cup \{x\}, A)$, a directed positional graph is a directed subgraph of $G$ such that

- If $(x, y_1) \in A$, $y_1, y_2 \in V_1$, and $y_3 \in V_2$, then
  $$(x, y_2) \in A, (y_3, x) \notin A$$

- If $(x, y_1) \in A$, $y_1, y_2 \in V_2$, and $y_3 \in V_1$, then
  $$(x, y_2) \in A, (y_3, x) \notin A$$

Having defined the various classes of subgraphs, we need to specify how they can be assembled:

**Definition 4.17** A graph $G = (N, A)$ is the sum of a set of graphs $G_i$ iff each $G_i$ is a subgraph of $G$ and

$$N = \bigcup N_i = \emptyset, A = \bigcup A_i = \emptyset$$
Figure 4.18: Contour graphs for surfaces in Figure 4.2.
Pulling all of this together, we get the class of connected figure/ground graphs, which will intuitively be isomorphic to the extension of *inside* for connected sets of edges:

**Definition 4.18** Given a structure $M$ in $L_{fg}$, a connected figure/ground graph for a surface is a directed graph $G = (N, A)$ with the following properties:

1. The set $N$ of nodes of the graph is a set of edges in some contour graph for the surface;
2. $G$ is isomorphic to the sum of a directed $C_n$ which is a directed subgraph of a contour graph in $M$, and the directed positional subgraphs for each edge in $N$.

**Example:** A connected figure/ground graph for the surface in Figure 4.2(a) is shown in Figure 4.19(a). The graph in (b) is a directed $C_n$ subgraph of the contour graph for the surface. The remaining graphs are the directed positional graphs for each edge in the surface. It is easy to see that the graph in (a) is indeed the sum of the graphs in (b) and (c).

There is another connected figure/ground graph for the surface in Figure 4.2(a), shown in Figure 4.20(a). As before, the graph in (b) is a directed $C_n$ subgraph of the contour graph for the surface. The remaining graphs are the directed positional graphs for each edge in the surface. It is easy to see that the graph in (a) is indeed the sum of the graphs in (b) and (c).

Note that since there are only two possible directed subgraphs of a $C_n$, the connected figure/ground graphs in Figure 4.19(a) and Figure 4.20(a) are the only two possible ones for the surface in Figure 4.2(a). □

**Example:** A connected figure/ground graph for the surface in Figure 4.2(b) are shown in Figure 4.21(a). The graph in (b) is a directed $C_n$ subgraph of the contour graph for the surface. The remaining graphs are the directed positional graphs for each edge in the surface. It is easy to see that the graph in (a) is indeed the sum of the graphs in (b),(c),(d),(e),(f), and (g).

The connected figure/ground graph for the surface in Figure 4.2(c) is isomorphic to the union of the graphs in Figure 4.19(a) and Figure 4.21(b), since the surface contains two disjoint cycles of edges.

□

Finally, we need to specify the classes of structures which characterize the figure/ground assignments for an entire surface. The motivation for the following two classes is intuitively to capture the extensions of both *inside* and *outside* for a surface.

Notice that because a connected figure/ground graph for a surface must contain a directed $C_n$ as a subgraph, and there are only two possible directed $C_n$ subgraphs for a cycle of edges, there are at most two possible connected figure/ground graphs for a surface. One of our intuitions for figure/ground assignments, however, is that every surface has a unique assignment. We must therefore specify a class of structures which determines this unique figure/ground assignment.
Figure 4.19: Connected figure/ground graph for the surface in Figure 4.2(a).

Figure 4.20: Alternative connected figure/ground graph for the surface in Figure 4.2(a).
Figure 4.21: Connected figure/ground graph for the surface in Figure 4.2(b).
Figure 4.22: Alternative connected figure/ground graph for the surface in Figure 4.2(b).
Definition 4.19 Given a structure $\mathcal{M} \in \mathcal{E}_{fg}$ and a connected figure/ground graph $G$, then $G$ is a contiguous figure/ground graph iff one of the following holds:

- If the edges in $G$ are outer edges, then $G$ contains a subgraph isomorphic to a connected subgraph of the contiguous graph for the edges in $G$.
- If the edges in $G$ are not outer edges, then $G$ contains a subgraph isomorphic to the inverse of a connected subgraph of the contiguous graph for the edges in $G$.

Example: Consider the surface in Figure 4.2(c). The graphs in Figure 4.23(a) are the two contiguous figure/ground graphs corresponding to the two sets of connected edges in the surface. The connected subgraphs of the contiguous graphs that correspond to the subgraphs of the figure/ground graphs are in 4.23(b).

Example: Consider the surface in Figure 4.2(a). The two possible connected figure/ground graphs for this surface are found in Figures 4.19 and 4.20. The contiguous graph for this surface is found in Figure 4.4(a). Notice that only the graph in Figure 4.19 contains a subgraph isomorphic to the contiguous graph for the surface. Therefore, it is the surface figure/ground graph, and the one in Figure 4.20 is not the surface figure/ground graph.

Consider the surface in Figure 4.2(b). The two possible connected figure/ground graphs for this surface are found in Figures 4.21 and 4.22. The contiguous graph for this surface is found in Figure...
Figure 4.24: Inverse figure/ground graph for surface in Figure 4.2(b).

Notice that only the graph in Figure 4.21 contains a subgraph isomorphic to the contiguous graph for the surface. Therefore, it is the surface figure/ground graph, and the one in Figure 4.22 is not the surface figure/ground graph.

Consider the surface in Figure 4.2(c). In particular, consider the edges $e_4, e_5, e_6, e_7, e_8$. The two possible connected figure/ground graphs for these edges are isomorphic to the graphs found in Figures 4.21(a) and 4.22(a). The contiguous graph for the hole edges in this surface is found in Figure 4.4(b). Notice that only the graph in Figure 4.22(a) contains a subgraph isomorphic to the contiguous graph for the surface. Therefore, it is a subgraph of the surface figure/ground graph, and the one in Figure 4.21(a) is not a subgraph of the surface figure/ground graph.

\[ \square \]

**Definition 4.20** Given a structure $\mathcal{M}$ in $\mathcal{L}_{fg}$, a surface figure/ground graph for a surface is a directed graph $G = (N, A)$ with the following properties:

1. The set $N$ of nodes of the graph is a set of edges in the surface;
2. $G$ is isomorphic to the sum of the contiguous figure/ground graphs for each disjoint cycle of edges in the surface, and the directed positional subgraphs for each edge in $N$.

All of the classes of figure/ground graphs to this point have intuitively been characterizing the extension of *inside*. We now turn to the class of graphs which characterize the extension of *outside*.

**Definition 4.21** Given a structure $\mathcal{M} \in \mathcal{L}_{fg}$, and a surface figure/ground graph $G = (N, A)$ for a surface $\mathcal{s}$, the inverse figure/ground graph $G' = (N, A')$ for $\mathcal{s}$ is the following graph:

- If $(e_1, e_2) \in \text{indet}$, then $(e_1, e_2) \in A'$ iff $(e_1, e_2) \in A$
- If $(e_1, e_2) \notin \text{indet}$, then $(e_1, e_2) \in A'$ iff $(e_1, e_2) \notin A$

**Example:** Consider the surface in Figure 4.2(a). In this surface, there are no indeterminate edges, so the inverse figure/ground graph is isomorphic to the inverse of the surface figure/ground graph. $\square$
Example: Consider the surface in Figure 4.2(b). In this surface, we have

\[ \langle e_1, e_3 \rangle, \langle e_2, e_4 \rangle \in \text{indet} \]

Consider the surface figure/ground graph from Figure 4.21. The inverse figure/ground graph for the surface is in Figure 4.24. Except for the arcs \( \langle e_1, e_3 \rangle, \langle e_2, e_4 \rangle \), these two graphs are isomorphic, while these two arcs are inverted between the two graphs. \( \square \)

Having defined all of the basic structures for the various relations, we can now put them together to give a characterization of all structures which we will show to be models of \( T_{fg} \).

**Definition 4.22** Let \( \mathcal{M}_i^{fg} \) be a set of structures with the following properties: For all structures \( \mathcal{M} \in \mathcal{M}_i^{fg} \),

1. \( \mathcal{M} \) is an extension of a structure in \( \mathcal{M}_i^{scene} \);
2. \( \mathcal{M} \) is an extension of a structure in \( \mathcal{M}_i^{polygon} \);
3. The extension of the relation meet in \( \mathcal{M} \) for edges in the same surface is isomorphic to a contour graph; that is, there exists an arc \( \langle e_1, e_2 \rangle \) in a contour graph iff there exists \( v \in \mathcal{M} \) such that \( \langle e_1, e_2, v \rangle \in \text{meet} \)
4. \( \langle e_1, e_2 \rangle \in \text{connected} \) iff the nodes for the edges are in the same contour graph in \( \mathcal{M} \).
5. The set of edges which are part of the same surface in \( \mathcal{M} \) is partitioned into sets which are isomorphic to contour graphs.
6. For each surface \( s \in \mathcal{M} \), there exists a unique contour graph such that an edge \( e \in \mathcal{M} \) is an element of the graph iff

\[ \langle e \rangle \in \text{outer} \text{ and } \langle e, s \rangle \in \text{part} \]
7. All edges which are part of the same surface in \( \mathcal{M} \) are members of two disjoint sets - the set of connected outer edges and the set of hole edges in the surface.
8. The extension of inside is isomorphic to a set of surface figure/ground graphs, such that for each surface in \( \mathcal{M} \) there exists a unique surface figure/ground graph in the set. That is, there exists a directed arc between two edges \( e_1, e_2 \) in a surface figure/ground graph iff

\[ \langle e_1, e_2 \rangle \in \text{inside} \]
9. The extension of outside is isomorphic to a set of inverse figure/ground graphs for each surface figure/ground graph in \( \mathcal{M} \), such that for each surface in \( \mathcal{M} \) there exists a unique inverse figure/ground graph in the set. That is, there exists a directed arc between two edges \( e_1, e_2 \) in an inverse figure/ground graph iff

\[ \langle e_1, e_2 \rangle \in \text{outside} \]

Conditions 1 and 2 specify the relationship between \( \mathcal{M}_i^{fg} \) and the previously specified classes of structures. Conditions 3-7 characterize the extensions of the relations \( \text{meet}, \text{connected}, \text{and outer} \) for structures in \( \mathcal{M}_i^{fg} \). Conditions 8 and 9 characterize the extensions of \( \text{inside} \) and \( \text{outside} \) for structures in \( \mathcal{M}_i^{fg} \).

It should be noted that none of the unintuitive figure/ground assignments in Figures 4.17 and 4.16 correspond to structures in \( \mathcal{M}_i^{fg} \). The extension \( \text{inside} \) in each case is not isomorphic to the surface figure/ground graph for the set of edges.
4.3.2 Characterization Theorems

Given the definition of $\mathcal{M}_i^g$, we can prove several theorems about the structures in this class.

We must first establish that the class of structures in $\mathcal{M}_i^g$ exists and is nonempty. (Without this result, the Satisfiability Theorem presented later in this chapter would be rather meaningless.)

**Theorem 4.2 Existence Theorem for $\mathcal{M}_i^g$**

*The class of structures $\mathcal{M}_i^g$ exists and is nonempty.*

**Proof:** We need to show that for each surface, each of the following graphs exist:

1. Contour graphs
   - This is trivial, since contour graphs are isomorphic to $C_n$ graphs, and the class of $C_n$ graphs is nonempty.

2. Butterfly graphs
   - The existence of this class of graphs follows from the existence of Hilbert's model of Euclidean geometry.

3. Figure/ground graphs
   - A figure/ground graph is the union of a set of arbitrary contour and butterfly graphs. Since each of these classes of graphs is nonempty, their union is also nonempty.

4. Inverse figure/ground graphs
   - An inverse figure/ground graph is the union of a set of arbitrary contour and butterfly graphs. Since each of these classes of graphs is nonempty, their union is also nonempty. Since structures in $\mathcal{M}_i^g$ are the union of structures in the above classes, and these classes are nonempty, we also have that $\mathcal{M}_i^g$ is nonempty.

$\square$

The following theorem characterizes the extensions of *inside* and *outside* between any pair of structures in $\mathcal{M}_i^g$:

**Theorem 4.3** If $\mathcal{M}, \mathcal{M}'$ are structures in $\mathcal{M}_i^g$ which have the same domain and which agree on the extensions of all polygonal relations, then they have isomorphic inside and outside graphs, i.e., they agree on the extensions of the relations *inside* and *outside*.

**Proof:** Suppose there exist two structures $\mathcal{M}_1, \mathcal{M}_2$ in $\mathcal{M}_i^g$ which have the same domain and which agree on the extensions of all polygonal relations, yet they disagree on the extensions of the relations *inside* and *outside*.

Then $\mathcal{M}_1$ and $\mathcal{M}_2$ have isomorphic contour graphs, butterfly graphs, and contiguous graphs. However, the figure/ground graphs are the union of these graphs, so that the figure/ground graphs in these two structures must also be isomorphic. Since the extension of *inside* is isomorphic to the set of figure/ground graphs, we have a contradiction.

A similar argument holds for inverse figure/ground graphs and the extension of *outside*. $\square$

Finally, the next two results characterize the decomposition of structures in $\mathcal{M}_i^g$.

**Theorem 4.4** The union of two structures in $\mathcal{M}_i^g$ is also a structure in $\mathcal{M}_i^g$. 


Proof: By the Depicted Kernel Theorem, all models of $T_{scene}$ are equivalent to the union of disjoint surfaces, so that the union of two models of $T_{scene}$ is also a model of $T_{scene}$.

Each structure in $\mathcal{M}_f^g$ is an extension of a model of $T_{scene}$. Further, each surface has a unique figure/ground graph and inverse figure/ground graph, and these graphs are disjoint for distinct surfaces. □

Corollary 4.1 A structure in $\mathcal{M}_f^g$ is minimal iff it contains a unique surface.

Proof: This is easy to see from the Depicted Kernel Theorem, since all models of $T_{scene}$ consist of disjoint surfaces. □

4.3.3 Presenting $T_{fg}$

The problem with which we are faced is to determine the extensions of inside and outside, given a surface with a complete assignment of alignment and positional relations. In this section, we introduce the axioms of $T_{fg}$ and show how they can be used to provide figure/ground assignments, how they intuitively satisfy the conditions of $\mathcal{M}_f^g$, and how they eliminate anomalous figure-ground assignments (which are unintended structures) that are not in $\mathcal{M}_f^g$.

The axioms of $T_{fg}$ can be found in Figures 4.25 and 4.26. Axioms 4.1 - 4.3 specify the meets relation, so that surfaces are isomorphic to contour graphs. Axiom 4.4 specifies the connected relation, so that edges in the same contour graph are connected. Notice that this is a second-order axiom, because connectedness for graphs is not first-order definable, even if we restrict ourselves to finite graphs ([Fagin 90]).

Axiom 4.14 guarantees the existence of outer edges, since for every edge in the surface there exists an outer edge which is contiguous with it. Axiom 4.5 constrains all outer edges to be connected, so that there is at most one set of outer edges in any surface. Axioms 4.6 and 4.7 can be considered to be closure axioms for the polygon relations – all edges in the same surface must have some relative alignment and all edges in the same surface must either be on the same side, opposite side, or indeterminate with each edge in the surface.

The heart of $T_{fg}$ is the set of axioms 4.8 - 4.13. Given the extensions of the polygon relations, these axioms specify the figure/ground assignment for any two connected edges in the same surface.

---

Example: Consider the surface represented in Figure 4.2(a). There exists a structure in $\mathcal{M}_f^g$ satisfying $T_{fg}$ such that \{e_1, e_2, e_3, v_1, v_2, v_3\} $\subseteq M$ and

$$\langle e_1, e_2, v_1 \rangle \in \text{meet}, \langle e_2, e_3, v_2 \rangle \in \text{meet}, \langle e_1, e_3, e_2 \rangle \in \text{opposite}$$

By axiom 4.10 we have

$$\langle e_1, e_2 \rangle \in \text{inside}, \langle e_1, e_2 \rangle \notin \text{outside}, \langle e_2, e_1 \rangle \notin \text{inside}, \langle e_2, e_1 \rangle \in \text{outside}$$
Axioms for meet

Two edges meet at a vertex iff they are distinct and the vertex is part of both edges.

\[(\forall e, e', v) \text{ meet}(e, e', v) \equiv \text{edge}(e) \land \text{edge}(e') \land \text{point}(v) \land \text{part}(v, e) \land \text{part}(v, e') \land e \neq e'\] (4.1)

Every edge meets two distinct edges.

\[(\forall e)(\exists e', e'', v, v') \text{ meet}(e, e', v) \land \text{meet}(e'', e, v') \land e' \neq e''\] (4.2)

Two edges meet at a unique vertex.

\[(\forall e, e', v, v') \text{ meet}(e, e', v) \land \text{meet}(e, e', v') \supset v = v'\] (4.3)

Axioms for outer edges:

Two edges are connected iff there exists a finite sequence of edges between them such that consecutive edges in the sequence meet at some vertex. (Note that this is a second-order sentence.)

\[\exists (\forall x, y) \text{ connected}(x, y) \equiv\]

\[\exists (\forall P) \left[\left[\left(\exists e_1, e_2, v \text{ meet}(e_1, e_2, v) \supset P(e_1, e_2)\right)\right] \land \left[\left(\exists e_1, e_2, e_3, v \text{ meet}(e_1, e_2, v) \supset P(e_1, e_2)\right) \land \text{meet}(e_2, e_3, v) \supset P(e_1, e_3)\right]\right] \supset P(x, y)\] (4.4)

All outer edges are connected to each other.

\[\exists (\forall e, e') \text{ outer}(e) \supset (\text{outer}(e') \equiv \text{connected}(e, e'))\] (4.5)

Axioms for polygon relations:

Every two edges in the same surface have an alignment relation.

\[(\forall e, e', s) \text{ edge}(e) \land \text{edge}(e') \land \text{part}(e, s) \land \text{part}(e', s) \equiv (\text{parallel}(e, e') \lor (\exists a) \theta(a, e, e'))\] (4.6)

Closure over the positional relations.

\[(\forall e_1, e_2, e_3) \left(\text{sameside}(e_1, e_2, e_3) \lor \text{opposite}(e_1, e_2, e_3)\right)\]

\[\equiv (\exists s) \left(\text{edge}(e_1) \land \text{edge}(e_2) \land \text{edge}(e_3) \land \text{part}(e_1, s) \land \text{part}(e_2, s) \land \text{part}(e_3, s)\right)\] (4.7)

Figure 4.25: $T_f$: The figure/ground axioms.
Axioms for figure/ground assignments:

\[(\forall e_1, e_2) \text{inside}(e_1, e_2) \supset (\exists s) \text{edge}(e_1) \land \text{edge}(e_2) \land \text{part}(e_1, s) \land \text{part}(e_2, s)\]  \hspace{1cm} (4.8)

\[(\forall e_1, e_2) \text{outside}(e_1, e_2) \supset (\exists s) \text{edge}(e_1) \land \text{edge}(e_2) \land \text{part}(e_1, s) \land \text{part}(e_2, s)\]  \hspace{1cm} (4.9)

\[(\forall e, e', v) \text{meet}(e, e', v) \supset (\text{inside}(e, e') \equiv \neg \text{inside}(e', e))\]  \hspace{1cm} (4.10)

\[(\forall e, e', e'', s) \text{sameside}(e', e'', e) \supset (\text{inside}(e, e') \equiv \text{inside}(e, e''))\]  \hspace{1cm} (4.11)

\[(\forall e, e', e'', s) \text{opposite}(e', e'', e) \supset (\text{inside}(e, e') \equiv \neg \text{inside}(e, e''))\]  \hspace{1cm} (4.12)

\[(\forall e, e') \text{indet}(e, e') \equiv (\text{inside}(e, e') \equiv \text{outside}(e, e'))\]  \hspace{1cm} (4.13)

\[(\forall e, s) \text{edge}(e) \land \text{part}(e, s) \supset (\exists e') \text{part}(e', s) \land \text{contiguous}(e, e') \land \text{outer}(e') \land \text{convex}(e, e')\]  \hspace{1cm} (4.14)

\[(\forall e, s) \neg \text{outer}(e) \land \text{part}(e, s) \supset (\exists e') \text{part}(e', s) \land \text{contiguous}(e, e') \land \text{connected}(e, e') \land \neg \text{convex}(e, e')\]  \hspace{1cm} (4.15)

Figure 4.26: \(T_{fg}\): The figure/ground axioms (cont.)
or
\[(e_1, e_2) \notin \text{inside}, (e_1, e_2) \in \text{outside}, (e_2, e_1) \in \text{inside}, (e_2, e_1) \notin \text{outside}\]

Since \((e_2, e_3, \nu_2) \in \text{meet}\), by axiom 4.10 we have
\[(e_2, e_3) \in \text{inside}, (e_2, e_3) \notin \text{outside}, (e_3, e_2) \notin \text{inside}, (e_3, e_2) \in \text{outside}\]
or
\[(e_2, e_3) \notin \text{inside}, (e_2, e_3) \in \text{outside}, (e_3, e_2) \in \text{inside}, (e_3, e_2) \notin \text{outside}\]

Since \((e_1, e_3, e_2) \in \text{opposite}\), by axiom 4.12 we must have
\[(e_2, e_1) \in \text{inside} \iff (e_2, e_3) \notin \text{inside}\]

so that all models will either satisfy
\[(e_1, e_2) \in \text{inside}, (e_1, e_2) \notin \text{outside}, (e_2, e_1) \notin \text{inside}, (e_2, e_1) \in \text{outside}\]
\[(e_2, e_3) \in \text{inside}, (e_2, e_3) \notin \text{outside}, (e_3, e_2) \notin \text{inside}, (e_3, e_2) \in \text{outside}\]
or
\[(e_1, e_2) \notin \text{inside}, (e_1, e_2) \in \text{outside}, (e_2, e_1) \in \text{inside}, (e_2, e_1) \notin \text{outside}\]
\[(e_2, e_3) \notin \text{inside}, (e_2, e_3) \in \text{outside}, (e_3, e_2) \in \text{inside}, (e_3, e_2) \notin \text{outside}\]

\(\square\)

The final two axioms (4.14 and 4.15) determine the correct figure/ground assignment for outer and hole edges. For every outer edge in the surface, axiom 4.14 asserts the existence of an outer edge which is contiguous with it. Given this edge, the figure/ground assignment is determined by the relative alignment of the outer contiguous edge. Similarly, for every hole edge, axiom 4.15 asserts the existence of a connected hole edge which is contiguous with it. Given this edge, the figure/ground assignment is determined by the relative alignment of the non-outer contiguous edge. These two axioms serve to disambiguate the possible figure/ground assignments which are consistent with axioms 4.8 - 4.13. If the surface has \(n\) different sets of closed contours, then the surface will have at most \(2^n\) complete figure ground assignments, since each set of connected edges has two possible figure-ground assignments that are consistent with axioms 4.8 - 4.13. However, our intuitions require that every surface have a unique figure/ground assignment. This assignment is fixed by the conditions in Axioms 4.14 and 4.15, which uniquely determine the figure/ground assignment using the relative alignment the edges. To enhance the readability of these axioms, we introduce the following defined relation:

**Definition 4.23**

\[(\forall e_1, e_2) \ \text{convex}(e_1, e_2) \ \overset{\text{def}}{=} (\text{inside}(e_1, e_2) \equiv (\text{parallel}(e_1, e_2) \vee (\exists a) \theta(a, e, e') \wedge a < \pi))\]
We can illustrate Axiom 4.14 by considering how it can be used to eliminate anomalous figure ground assignments.

**Example:** Consider the two possible figure ground assignment for the surfaces in Figure 4.16. In (a), the edges form the boundary of a hole, yet there is no boundary to the surface. In (b), although the edges e₅ through e₈ are the boundary of a hole, the assignment indicates that there is no hole. Axiom 14 is simply the definition of a relation that we will use to eliminate these anomalous figure ground assignments.

How does this eliminate the anomalous figure ground assignments? Consider Figure 4.16a: this assignment violates the axiom since there does not exist an edge e which is both contiguous with e₂ and which satisfies (e, e₂) \( \in \) inside iff \( (a, e, e₂) \in \theta \) and \( a < \pi \). In Figure 4.16b, there is no outer edge that is contiguous with edge e₄ which also satisfies (e, e₄) \( \in \) inside iff \( (a, e, e₄) \in \theta \) and \( a < \pi \). Thus the two figure ground assignments illustrated in Figure 4.16 are not feasible. □

We will now show that the axioms in \( T_{fg} \) are satisfied by the class of structures in \( \mathcal{M}^{fg} \), and then characterize the models of \( T_{fg} \).

### 4.3.4 Satisfiability of \( T_{fg} \)

We can formalize the discussion of the preceding section by proving the satisfiability of the axioms of \( T_{fg} \): that is, we will show that any structure which satisfies the definition of \( \mathcal{M}^{fg} \) also satisfies the axioms of \( T_{scene} \cup T_{fg} \).

**Theorem 4.5** Any structure \( \mathcal{M} \in \mathcal{M}^{fg} \) is a model of \( T_{scene} \cup T_{fg} \).

**Proof:** Since all structures in \( \mathcal{M}^{fg} \) are extensions of structures in \( \mathcal{M}_{scene} \), by the Axiomatizability Theorem for \( T_{kernel} \) they will satisfy \( T_{scene} \).

Next, consider a structure \( \mathcal{M} \) which satisfies conditions 3-5 in the definition of \( \mathcal{M}^{fg} \), so that the extension of meet is isomorphic to a contour graph.

\[ \mathcal{M} \models (\forall e₁, e₂, v) \text{ meet}(e₁, e₂, v) \equiv \text{edge}(e₁) \land \text{edge}(e₂) \land \text{point}(v) \land \text{part}(v, e₁) \land \text{part}(v, e₂) \land e₁ \neq e₂ \]

is equivalent to saying that for any variable assignment \( \sigma \),

\[ \mathcal{M}, \sigma \models \text{meet}(e₁, e₂, v) \]

iff

\[ \mathcal{M}, \sigma \models \text{edge}(e₁) \land \text{edge}(e₂) \land \text{point}(v) \land \text{part}(v, e₁) \land \text{part}(v, e₂) \land e₁ \neq e₂ \]

which is equivalent to

\[ (e₁, e₂, v) \in \text{meet} \]

iff

\[ (e₁) \in \text{edge}, (e₂) \in \text{edge}, (v) \in \text{point}, \quad (v, e₁) \in \text{part}, (v, e₂) \in \text{part} \]

for each tuple of distinct elements \( e₁, e₂, v \). This follows from the construction of the contour graph for a surface, in which there is an arc in the graph iff there is a common point which is part of two edges in the surface. Therefore Axiom 4.1 is satisfied.
iff for any element \( e_1 \in M \) there exist distinct elements \( e_2, e_3, v_1, v_2 \in M \) such that
\[
\langle e_1, e_2, v_1 \rangle, \langle e_1, e_2, v_2 \rangle \in \text{meet}
\]
Similarly
\[
M \models (\forall e, e', v, v') \text{ meet}(e, e', v) \land \text{meet}(e', e', v') \supset v = v'
\]
iff for any elements \( e_1, e_2 \in M \) there exists a unique element \( v \in M \) such that
\[
\langle e_1, e_2, v \rangle \in \text{meet}
\]
Both of these conditions follow because in any contour graph which is a substructure of \( M \) and which is isomorphic to the extension of \( \text{meet} \), every node is adjacent to exactly two other nodes. Therefore, \( M \) satisfies Axioms 4.2 and 4.3.
\( M \) satisfies axiom 4.4 (This follows from a general result on connected graphs from general graph theory [Doets 96]).
Since there exists a unique set of outer edges in \( M \) and all outer edges are connected,
\[
M \models (\forall e_1, e_2) \text{outer}(e_1) \land \text{outer}(e_2) \supset \text{connected}(e_1, e_2)
\]
Further, no hole edge is connected to any outer edge, so that we have
\[
M \models (\forall e_1, e_2) \text{outer}(e_1) \land \neg \text{outer}(e_2) \supset \neg \text{connected}(e_1, e_2)
\]
Thus, \( M \) satisfies axiom 4.5.
\( M \) satisfies axioms 4.6 - 4.7, since by the definition of \( M_{\text{polygon}}^i \) and condition (2) in the definition of \( M_{\text{fig}}^i \), every pair of edges in a surface has a relative alignment, and every triple of edges in a surface are elements in a butterfly graph which is a substructure of \( M \).
\[
M \models (\forall e_1, e_2) \text{inside}(e_1, e_2) \supset (\exists s) \text{edge}(e_1) \land \text{edge}(e_2) \land \text{part}(e_1, s) \land \text{part}(e_2, s)
\]
iff for any elements \( e_1, e_2 \in M \) there exists an element \( s \in M \) such that
\[
\langle e_1, s \rangle \in \text{part}, \langle e_2, s \rangle \in \text{part}
\]
This follows from the construction of the connected figure/ground graph in which all nodes of the graph are edges which are part of the same surface. A similar argument shows that
\[
M \models (\forall e_1, e_2) \text{outside}(e_1, e_2) \supset (\exists s) \text{edge}(e_1) \land \text{edge}(e_2) \land \text{part}(e_1, s) \land \text{part}(e_2, s)
\]
and hence \( M \) satisfies axioms 4.8 and 4.9.
\( M \) satisfies axioms 4.10 - 4.13:
\[
M \models (\forall e_1, e_2, v) \text{meet}(e_1, e_2, v) \supset (\text{inside}(e_1, e_2) \equiv \neg \text{inside}(e_2, e_1))
\]
iff for any elements \( e_1, e_2, v \)
\[
\langle e_1, e_2, v \rangle \in \text{meet}
\]
implies
\[
\langle e_1, e_2 \rangle \in \text{inside} \iff \langle e_2, e_1 \rangle \notin \text{inside}
\]
Since the connected figure/ground graph contains a subgraph which is a directed \( C_n \) subgraph of a contour graph, we have
\( (e_1, e_2) \in \text{inside} \) or \( (e_2, e_1) \in \text{inside} \)

and if \( (e_1, e_2) \in \text{inside}, \) then \( (e_2, e_1) \notin \text{inside} \) (because the subgraph is a directed \( C_n \) and cannot contain 2-cycles). Therefore \( \mathcal{M} \) satisfies axiom 4.10.

\[
\mathcal{M} \models (\forall e_1, e_2, e_3) \text{sameside}(e_1, e_2, e_3) \supset (\text{inside}(e_1, e_2) \equiv \text{inside}(e_1, e_3))
\]

iff for any elements \( e_1, e_2, e_3 \in M, \) if

\[ (e_1, e_2, e_3) \in \text{sameside} \]

then \( (e_1, e_2) \in \text{inside} \) iff \( (e_2, e_1) \notin \text{inside} \)

This follows from the definition of \( \text{inside} \), since the connected figure/ground graph contains a subgraph which is isomorphic to a directed subgraph of a sameside graph (which is a claw graph), the arcs to all leaves of the claw will have the same direction. Therefore, \( \mathcal{M} \) satisfies axiom 4.11.

\[
\mathcal{M} \models (\forall e_1, e_2, e_3) \text{opposite}(e_1, e_2, e_3) \supset (\text{inside}(e_1, e_2) \equiv \neg \text{inside}(e_1, e_3))
\]

iff for any elements \( e_1, e_2, e_3 \in M, \) if

\[ (e_1, e_2, e_3) \in \text{opposite} \]

then \( (e_1, e_2) \in \text{inside} \) iff \( (e_2, e_1) \notin \text{inside} \)

This follows from the definition of \( \text{inside} \), since the connected figure/ground graph contains a subgraph which is isomorphic to the complement of an opposite graph.

\[
\mathcal{M} \models (\forall e_1, e_2) \text{indet}(e_1, e_2) \equiv (\text{inside}(e_1, e_2) \equiv \text{outside}(e_1, e_2))
\]

iff for any elements \( e_1, e_2 \in M, \) if

\[ (e_1, e_2) \in \text{indet}, \text{then} \ (e_1, e_2) \in \text{inside} \) iff \( (e_1, e_2) \in \text{outside} \)

and if

\[ (e_1, e_2) \notin \text{indet}, \text{then} \ (e_1, e_2) \notin \text{inside} \) iff \( (e_1, e_2) \notin \text{outside} \)

This follows from the definition of inverse figure/ground graphs, because the the subgraph of the \( \text{inside} \) graph containing indeterminate edges is isomorphic to the subgraph of the \( \text{outside} \) graph containing indeterminate edges, and the inverse of the subgraph of the \( \text{outside} \) graph containing non-indeterminate edges is isomorphic to the subgraph of the \( \text{inside} \) graph containing non-indeterminate edges. Therefore \( \mathcal{M} \) satisfies axiom 4.13.

\[
\mathcal{M} \models (\forall e_1, s) \text{outer}(e_1) \land \text{part}(e_1, s) \supset (\exists e_2) \text{contiguous}(e_1, e_2) \land \text{outer}(e_2) \land \text{convex}(e_1, e_2)
\]

iff for any outer edge \( e_1 \in M \) there exists an element \( e_2 \in M \) such that

\[ (e_1, e_2) \in \text{contiguous}, (e_2) \in \text{outer}, (e_1, e_2) \in \text{convex} \]

This follows from the definition of the surface figure/ground graph for a surface, since each connected figure/ground graph contains a subgraph isomorphic to the subgraph of the contiguous graph containing outer edges. Therefore \( \mathcal{M} \) satisfies axiom 4.14. A similar argument holds for axiom 4.15.

\[ \square \]
4.3.5 Axiomatizability Theorem for $T_{fg}$

The axiomatizability theorem for $T_{fg}$ demonstrates that we do not have any unintended models of the axioms which disagree with the intuitions characterized in the definition of the class of structures in $\mathcal{M}^{fg}_i$.

**Theorem 4.6** Any model of $T_{\text{scene}} \cup T_{fg}$ which is an extension of a structure in $\mathcal{M}^{\text{polygon}}_i$ is isomorphic to a structure in $\mathcal{M}^{fg}_i$.

We will prove this theorem by showing how every model of $T_{\text{scene}} \cup T_{fg}$ satisfies each of the conditions in the definition of $\mathcal{M}^{fg}_i$.

**Contour Graphs in Models of $T_{fg}$**

We will begin by considering how $T_{fg}$ axiomatizes the contour graphs for surfaces.

We first need the following lemma:

**Lemma 4.1** Only two edges meet at a vertex.

$$T_{\text{scene}} \cup T_{fg} \models (\forall e_1, e_2, v) \text{meet}(e_1, e_2, v) \land \text{meet}(e_1, e_3, v) \supset e_2 = e_3$$

**Proof:** Suppose

$$T_{\text{scene}} \cup T_{fg} \models (\exists e_1, e_2, e_3, v) \text{meet}(e_1, e_2, v) \land \text{meet}(e_1, e_3, v) \land (e_2 \neq e_3)$$

By axiom 4.1 of $T_{fg}$, we have

$$T_{\text{scene}} \cup T_{fg} \models (\exists e_1, e_2, e_3, v) \text{part}(v, e_1) \land \text{part}(v, e_2) \land \text{part}(v, e_3)$$

which violates $T_{\text{scene}}$, since a point can be part of at most two edges. □

**Lemma 4.2** In any model of $T_{\text{scene}}$ satisfying axioms 4.1 - 4.3 of $T_{fg}$ containing a unique surface with $n$ edges and with no hole edges, the extension of meet is isomorphic to a $C_n$ graph.

**Proof:** By induction on the set of edges in a model $\mathcal{M}$ of $T_{\text{scene}} \cup T_{fg}$.

Base case: $n=3$

It is easy to see that the only structure which satisfies axioms 4.1 - 4.4 satisfies:

$$\mathcal{M}, \sigma \models \text{meet}(e_1, e_2, v_1) \land \text{meet}(e_2, e_3, v_2) \land \text{meet}(e_3, e_1, v_3)$$

which is isomorphic to a $C_3$ graph.

Induction step: Suppose that we have a set of $n$ edges satisfying

$$\mathcal{M}, \sigma \models \text{meet}(e_1, e_2, v_1) \land ... \land \text{meet}(e_{n-1}, e_n, v_{n-1})$$

By axiom 4.2, the edge $\sigma(e_n)$ must meet some other edge in this set. Such an edge cannot be $\sigma(e_i)$ for $i > 1$, otherwise this edge would meet three distinct edges, contradicting the previous lemma. Therefore, the only possible edge that can meet $\sigma(e_n)$ is $\sigma(e_1)$, so that

$$\mathcal{M}, \sigma \models \text{meet}(e_n, e_1, v_n)$$

and hence the set of $n$ edges is isomorphic to a $C_n$ graph. □

\(^3\)Note that this is the base case, since by $T_{\text{scene}}$ all surfaces must contain at least three edges.
Lemma 4.3 In any model of $T_{scene}$ satisfying axioms 4.1 - 4.3 of $T_{fg}$, the extension of meet is isomorphic to a set of $C_n$ graphs.

Proof: Induction on the number of disjoint cycles.

Base case ($n = 1$): This follows from the preceding lemma.

Induction step: Suppose we have a model of $T_{scene}$ satisfying axioms 4.1 - 4.3 of $T_{fg}$ consisting of $n$ cycles of edges which are part of the same surface.

Now adjoin to this model a distinct cycle of edges which are part of the same surface. Since we are adjoining a cycle of edges, axioms 4.1 - 4.3 are already satisfied for the new elements.

Thus the union of this model with a new disjoint cycle of edges which are part of the same surface is also a model of $T_{scene}$ satisfying axioms 4.1 - 4.3 of $T_{fg}$. □

Next we consider how $T_{fg}$ axiomatizes connected sets of edges:

Lemma 4.4 In any model $M$ of $T_{scene} \cup T_{fg}$,

$$\langle e_1, e_m \rangle \in \text{connected}$$

iff $e_1$ and $e_m$ are elements in the same $C_n$ graph.

Proof: We will first prove that if $e_1$ and $e_m$ are elements in the same $C_n$ graph, then

$$\langle e_1, e_m \rangle \in \text{connected}$$

In particular, we will show that if there exists a sequence of adjacent edges $[e_1, e_2, ..., e_m]$ from edge $e_1$ to edge $e_m$, then the edges are connected.

Assume that we have this sequence of edges, so that

$$\langle e_1, e_2, v_2 \rangle, ..., \langle e_{m-1}, e_m, v_m \rangle \in \text{meet}$$

Let $P$ be any relation satisfying the antecedent of the definition of connected in axiom 4.4. We must show that for all $k < m$, we have $\langle e_1, e_{k-1} \rangle \in \text{connected}$ in a sequence of edges. Obviously, the result holds for $k = 1$, since $\langle e_1, e_2, v_2 \rangle \in \text{meet}$. For the induction hypothesis, suppose that $\langle e_1, e_{j-1} \rangle \in \text{connected}$. If $j + 1 \neq n$, then it follows that $P(e_1, e_{j-1})$ (by definition of connected), and since we have a sequence of edges, we have $\langle e_{j-1}, e_{j-2}, v_{j-2} \rangle \in \text{meet}$, and hence $P(e_1, e_{j-2})$. So for all $k < n$, we have $\langle e_1, e_{k-1} \rangle \in \text{connected}$, by axiom 4.4.

For the converse direction, we can recursively define the set of all sequences of edges from $e_1$ to any edge that it is connected to in $M$. Consider the following sets of sequences:

$$A_1 = \{ [e, e_j] : \langle e_1, e \rangle \in \text{connected}, \langle e, e_j, v_j \rangle \in \text{meet} \}$$

$$A_{n-1} = \{ [e, e_n, ..., e_2, e_j] : \langle e_1, e \rangle \in \text{connected}, \langle e, e_n, v_n \rangle \in \text{meet}, \text{ and } [e_n, ..., e_2, e_j] \in A_n \}$$

We need to show that if $\langle e_1, e_j \rangle \in \text{connected}$, there must be a sequence of adjacent edges from $e_1$ to $e_j$ in some $A_m$.

Let $k$ be the least number such that $A_{k-1}$ is empty. If $k$ is 0, then there is no $e$ such that $\langle e_1, e \rangle \in \text{connected}$ and $\langle e, e_j, v_j \rangle \in \text{meet}$; by the first conjunct in the antecedent of the definition of connected in axiom 4.4, the only possibility is $\langle e_1, e_j, v_j \rangle \in \text{meet}$, so that there is a sequence of edges from $e_1$ to $e_j$. Suppose then that $k > 0$; then $A_k$ is not empty, and for all sequences
Inside, Outside, and Figure/Ground Graphs

We now move on to show how the extension of inside and outside in models of $T_{\text{scene}} \cup T_{fg}$ is isomorphic to sets of surface figure/ground graphs. The lemmas parallel the introduction of the various classes of structures – we first consider connected figure/ground graphs, then contiguous figure/ground graphs, and finally surface figure/ground graphs.

Lemma 4.5 In any model of $T_{\text{scene}} \cup T_{fg}$, the extension of inside for any connected set of edges is isomorphic to a connected figure/ground graph.

Proof: We first construct a graph $G = (N, A)$ such that $N$ is the set of edges in the surface, and $(\sigma(e_1), \sigma(e_2)) \in A$ iff

$$M, \sigma \models \text{inside}(\sigma(e_1), \sigma(e_2))$$

Suppose that we have a contour graph (connected set of $n$ edges):

$$M, \sigma \models \text{meet}(e_1, e_2, v_1) \land \cdots \land \text{meet}(e_{n-1}, e_n, v_{n-1}) \land \text{meet}(e_n, e_1, v_n)$$

$$\land \text{connected}(e_1, e_1) \land \cdots \land \text{connected}(e_n, e_n)$$

By axiom 4.10 $T_{fg}$, we get

$$M, \sigma \models \text{inside}(e_n, e_1) \equiv \neg \text{inside}(e_1, e_n)$$

and for all $1 < i < n$

$$M, \sigma \models \text{inside}(e_{i-1}, e_i) \equiv \neg \text{inside}(e_i, e_{i-1})$$

From this, we either get a subgraph such that $(\sigma(e_i), \sigma(e_{i-1})) \in A$ and $(\sigma(e_n), \sigma(e_1)) \in A$, or we get a subgraph such that $(\sigma(e_{i-1}), \sigma(e_i)) \in A$ and $(\sigma_1, \sigma_n)) \in A$. In either case, we have a subgraph which is isomorphic to a directed $C_n$ graph.

Since two edges which meet another edge are on opposite sides of this other edge, by axiom 4.12 $T_{fg}$ we have

$$M, \sigma \models \text{inside}(e_{n-1}, e_n) \equiv \neg \text{inside}(e_n, e_{n-1})$$

$$M, \sigma \models \text{inside}(e_n, e_{n-1}) \equiv \neg \text{inside}(e_n, e_1)$$

$$M, \sigma \models \text{inside}(e_1, e_{n-1}) \equiv \neg \text{inside}(e_1, e_n)$$

For all other edges, by axiom 4.7 $T_{fg}$ we have two cases:

1. $M, \sigma \models \text{sameside}(e_i, e_j, e_k)$

Axiom 4.11 $T_{fg}$ gives

$$M, \sigma \models \text{inside}(e_k, e_i) \equiv \text{inside}(e_k, e_j)$$

From this we get a subgraph such that either $(e_k, e_i), (e_k, e_j) \in A$ or $(e_i, e_k), (e_j, e_k) \in A$. In both cases, the subgraph is isomorphic to a directed positional graph with root node $e_k$. 

[ek, ..., e2, ej] ∈ A_k we have ⟨e1, ek⟩ ∈ connected. Since A_k−1 is nonempty, there is no edge e such that ⟨e1, e⟩ ∈ connected and ⟨e, ek, vk⟩ ∈ meet, for any sequence [ek, ..., e2, ej] ∈ A_k.

Hence, by second conjunct in the antecedent of the definition of connected in axiom 4.4, for any sequence [ek, ..., e2, ej] ∈ A_k, we can have ⟨e1, ek⟩ ∈ connected only if ⟨e1, ek, vk⟩ ∈ meet, so that we again have a sequence of edges from e1 to ej. □
2.

\[ M, \sigma \models \text{opposite}(e_i, e_j, e_k) \]

Axiom 4.12 \( T_{fg} \) gives

\[ M, \sigma \models \text{inside}(e_k, e_i) \equiv \neg \text{inside}(e_k, e_j) \]

From this, we get a subgraph such that either \((e_k, e_i) \in A\) and \((e_k, e_j) \notin A\) or \((e_k, e_j) \in A\) and \((e_k, e_i) \notin A\). In both cases, the subgraph is isomorphic to a directed positional graph with root node \( e_k \).

It is easy to see from the definition that \( G \) is isomorphic to a connected figure/ground graph. □

**Lemma 4.6** In any model of \( T_{\text{scene}} \cup T_{fg} \), the extension of outside for any connected set of edges is isomorphic to a connected figure/ground graph.

**Proof:** Suppose that we have a connected set of \( n \) edges:

\[ M, \sigma \models \text{meet}(e_1, e_2, v_1) \land \ldots \land \text{meet}(e_{n-1}, e_n, v_{n-1}) \land \text{meet}(e_n, e_1, v_n) \]

\[ \land \text{connected}(e_1, e_i) \land \ldots \land \text{connected}(e_n, e_n) \]

We can construct a connected figure/ground graph isomorphic to the extension of inside using the construction in the proof of the preceding lemma.

For any pair of edges \( \sigma(e_i), \sigma(e_j), 1 \leq i, j \leq n \) for some variable assignment \( \sigma \), we have two cases:

1. \[ M, \sigma \models \text{indet}(e_i, e_j) \]
   in which case we get \[ M, \sigma \models \text{outside}(e_k, e_i) \]
   iff \[ M, \sigma \models \text{inside}(e_i, e_j) \]

2. \[ M, \sigma \models \neg \text{indet}(e_i, e_j) \]
   in which case we get \[ M, \sigma \models \text{outside}(e_i, e_j) \]
   iff \[ M, \sigma \models \neg \text{inside}(e_i, e_j) \]

Since the extension of inside is isomorphic to a connected figure/ground graph, then in any model satisfying these conditions, the extension of outside is isomorphic to an inverse connected figure/ground graph. □

By the construction of the connected figure/ground graphs in the proofs of the preceding theorems, the following corollary is an easy consequence:

**Corollary 4.2** If \( M, M' \) are two models of \( T_{\text{scene}} \cup T_{fg} \) that have the same domain and that agree on the extensions of all relations in \( L_{\text{polygon}} \), then the extension of the relation inside for a set of connected edges in the two models is either equal or disjoint and the extension of the relation outside in the two models is either equal or disjoint.
Lemma 4.7 In any model of $T_{\text{scene}} \cup T_{fg}$, the extension of inside for any surface is isomorphic to a surface figure/ground graph.

Proof: Consider a set of connected figure/ground graphs $G_1, \ldots, G_n$ for a surface. By axiom 4.5, one of these graphs (say $G_1$) will contain the outer edges of the surface and the remaining graphs ($G_2, \ldots, G_n$) will contain the hole (non-outer) edges in the surface.

We will show that each of these graphs are contiguous figure/ground graphs.

Consider first a graph $G_2$ which contains hole edges. By axiom 4.15, for any edge $\sigma(e_1)$ in $G_2$, there exists a connected edge $\sigma(e_2)$ which is contiguous with $\sigma(e_2)$ and such that

$$M, \sigma \models \text{inside}(e_1, e_2) \equiv \neg(\text{parallel}(e_1, e_2) \lor (\theta(a, e_1, e_2) \land a < \pi))$$

so that in $G_2$ we have $(\sigma(e_1), \sigma(e_2)) \in A$ iff

$$M, \sigma \models \text{contiguous}(e_1, e_2) \land \neg(\text{parallel}(e_1, e_2) \lor (\theta(a, e_1, e_2) \land a < \pi))$$

Thus, $\sigma(e_1)$ and $\sigma(e_2)$ are elements of a subgraph of $G_2$ which is isomorphic to the inverse of the subgraph of the contiguous graph containing $\sigma(e_1)$ and $\sigma(e_2)$, and hence $G_2$ is a contiguous figure/ground graph.

Next consider the connected figure/ground graph $G_1$ which contains the outer edges. By axiom 4.14, for any edge $\sigma(e_1)$ in $G_1$, there exists a connected edge $\sigma(e_2)$ which is contiguous with $\sigma(e_2)$ and such that

$$M, \sigma \models \text{inside}(e_1, e_2) \equiv (\text{parallel}(e_1, e_2) \lor (\theta(a, e_1, e_2) \land a < \pi))$$

so that in $G_2$ we have $(\sigma(e_1), \sigma(e_2)) \in A$ iff

$$M, \sigma \models \text{contiguous}(e_1, e_2) \land (\text{parallel}(e_1, e_2) \lor (\theta(a, e_1, e_2) \land a < \pi))$$

Thus, $\sigma(e_1)$ and $\sigma(e_2)$ are elements of a subgraph of $G_2$ which is isomorphic to the subgraph of the contiguous graph containing $\sigma(e_1)$ and $\sigma(e_2)$, and hence $G_2$ is a contiguous figure/ground graph.

Finally, we need to show that the sum of these graphs is a surface figure/ground graph. Let $G(N, A)$ be a graph containing all of the connected figure/ground graphs $G_1, \ldots, G_n$.

Consider two outer edges $\sigma(e_1)$ and $\sigma(e_2)$, such that

$$M, \sigma \models \text{meet}(e_1, e_2, v)$$

By axiom 4.14, we know the figure/ground assignment for these two edges. For all other edges $\sigma(e_i)$, by axiom 4.7 $T_{fg}$ we have two cases:

1.

$$M, \sigma \models \text{sameside}(e_1, e_1, e_2)$$

Axiom 4.11 $T_{fg}$ gives

$$M, \sigma \models \text{inside}(e_2, e_1) \equiv \text{inside}(e_2, e_1)$$

From this we get a subgraph such that either $(\sigma(e_2), \sigma(e_1)), (\sigma(e_2), \sigma(e_1)) \in A$ or $(\sigma(e_1), \sigma(e_2)), (e_1, e_2) \in A$. In both cases, the subgraph is isomorphic to a directed positional graph with root node $\sigma(e_2)$.

2.

$$M, \sigma \models \text{opposite}(e_1, e_1, e_2)$$

Axiom 4.12 $T_{fg}$ gives

$$M, \sigma \models \text{inside}(e_2, e_1) \equiv \neg\text{inside}(e_2, e_1)$$

From this, we get a subgraph such that either $(\sigma(e_2), \sigma(e_1)) \in A$ and $(\sigma(e_2), \sigma(e_1)) \notin A$ or $(\sigma(e_2), \sigma(e_1)) \in A$ and $(\sigma(e_2), \sigma(e_1)) \notin A$. In both cases, the subgraph is isomorphic to a directed positional graph with root node $\sigma(e_2)$. 
It is easy to see from the definition that $G$ is isomorphic to a surface figure/ground graph. □

Lemma 4.8 In any model of $T_{scene} \cup T_{fg}$, the extension of outside for any surface is isomorphic to an inverse surface figure/ground graph.

Proof: We can construct a surface figure/ground graph isomorphic to the extension of inside using the construction in the proof of the preceding lemma.

For any pair of edges $\sigma(e_i), \sigma(e_j), 1 \leq i, j \leq n$ for some variable assignment $\sigma$, we have two cases:

1. 
   $M, \sigma \models \text{indet}(e_i, e_j)$
   in which case we get
   $M, \sigma \models \text{outside}(e_k, e_i)$
   iff
   $M, \sigma \models \text{inside}(e_i, e_j)$

2. 
   $M, \sigma \models \neg\text{indet}(e_i, e_j)$
   in which case we get
   $M, \sigma \models \text{outside}(e_i, e_j)$
   iff
   $M, \sigma \models \neg\text{inside}(e_i, e_j)$

Since the extension of inside is isomorphic to a surface figure/ground graph, then in any model satisfying these conditions, the extension of outside is isomorphic to an inverse surface figure/ground graph. □

Lemma 4.9 In any model of $T_{scene} \cup T_{fg}$, the extension of inside is isomorphic to a set of disjoint surface figure/ground graphs and the extension of outside is isomorphic to a set of disjoint inverse figure/ground graphs.

Proof: By axiom 4.13, we can see that every surface will have a nonempty extension of inside and outside, and hence, every surface will have a surface figure/ground graph and an inverse figure/ground graph.

By axioms 4.8 and 4.9, only edges which are part of the same surface can be elements of the same surface figure/ground graph, so the surface figure/ground graphs and inverse figure/ground graphs for distinct surfaces will be disjoint.

By the Axiomatizability Theorem for $T_{kernel}$, models of $T_{scene}$ consist of sets of surfaces. Thus, in any model of $T_{scene} \cup T_{fg}$, there will be a set of disjoint surface figure/ground graphs and a set of disjoint inverse figure/ground graphs. □

Using this result, it is easy to see the following two corollaries, which are structure theorems for models of $T_{scene} \cup T_{fg}$:

Corollary 4.3 The union of two models of $T_{scene} \cup T_{fg}$ is also a model of $T_{scene} \cup T_{fg}$.

Corollary 4.4 Minimal models in $T_{fg}$ are equivalent to unique surfaces.
Outer Edges in Models of $T_{fg}$

We next prove that models of $T_{scene} \cup T_{fg}$ satisfy the conditions in the definition of $\mathcal{M}^f$ dealing with the outer edges of a surface.

**Lemma 4.10** In any model of $T_{scene} \cup T_{fg}$, every surface has a unique set of outer edges.

**Proof:** By Axiom 4.14, for every surface there exists an outer edge. By Axiom 4.5, every edge which is connected to this edge is also an outer edge, so there cannot exist two disjoint sets of outer edges.  

**Lemma 4.11** If $\mathcal{M}, \mathcal{M}'$ are two models of $T_{scene} \cup T_{fg}$ that have the same domain and that agree on the extensions of all relations in $L_{polygon}$, and that agree on the extension of meet, then they agree on the extension of the relation outer.

**Proof:** Suppose that $\mathcal{M}, \mathcal{M}'$ are two models of $T_{scene} \cup T_{fg}$ that have the same domain and that agree on the extensions of all relations in $L_{polygon}$, but which do not agree on the extension of the relation outer.

Then for any variable assignment $\sigma$, there exists edges $\sigma(e_1), \sigma(e_2)$, such that

$$\mathcal{M}, \sigma \models \text{outer}(e_1) \land \neg \text{outer}(e_2)$$

$$\mathcal{M}', \sigma \models \neg \text{outer}(e_1) \land \text{outer}(e_2)$$

Since the models agree on the extensions of relations in $L_{polygon}$, by axiom 4.14 of $T_{fg}$, both models satisfy

$$\mathcal{M}, \sigma \models \text{inside}(e_1, e_2) \equiv (\text{parallel}(e_1, e_2) \lor (\theta(a, e_1, e_2) \land a < \pi))$$

$$\mathcal{M}', \sigma \models \text{inside}(e_1, e_2) \equiv (\text{parallel}(e_1, e_2) \lor (\theta(a, e_1, e_2) \land a < \pi))$$

By Corollary 4.2, both $\mathcal{M}$ and $\mathcal{M}'$ agree on the extension of inside for all edges that are in the cycles containing $\sigma(e_1)$ and $\sigma(e_2)$.

Now consider another edge $\sigma(e_3)$ which is connected to $\sigma(e_2)$ and which is an outer edge in $\mathcal{M}'$. By axiom 4.14 of $T_{fg}$, we have

$$\mathcal{M}', \sigma \models \text{inside}(e_2, e_3) \equiv (\text{parallel}(e_2, e_3) \lor (\theta(a, e_2, e_3) \land a < \pi))$$

However, since $\sigma(e_2)$ and $\sigma(e_3)$ are not outer edges in $\mathcal{M}$, we have

$$\mathcal{M}, \sigma \models \text{inside}(e_2, e_3) \equiv \neg (\text{parallel}(e_2, e_3) \lor (\theta(a, e_2, e_3) \land a < \pi))$$

which contradicts the previous claim that both $\mathcal{M}$ and $\mathcal{M}'$ agree on the extension of inside for all edges that are in the cycles containing $\sigma(e_1)$ and $\sigma(e_2)$.  

**Uniqueness of Figure/Ground Assignments**

Finally, we show that the axioms of $T_{fg}$ are powerful enough to prove that in models of $T_{scene} \cup T_{fg}$, each surface has a unique figure ground assignment.

**Theorem 4.7** If $\mathcal{M}, \mathcal{M}'$ are two models of $T_{scene} \cup T_{fg}$ that have the same domain and that agree on the extensions of all relations in $L_{polygon}$, then they agree on the extensions of the relations inside and outside.
Proof: Since the axioms in $T_{fg}$ deal only with edges in the same surface, we can restrict our proof to considering the edges in an arbitrary surface.

We have already shown that the extension of \textit{inside} within a model of $T_{fg}$ is isomorphic to a surface figure/ground graph. We therefore only need to show that for any surface this graph is unique.

The construction of the surface figure/ground graph is dependent on the extension of the \textit{outer} relation within the model $M$. By the Lemma 4.11, all models agree on the extension of \textit{outer}.

Therefore, all models of $T_{scene} \cup T_{fg}$ that have the same domain and that agree on the extensions of the relations $L_{\text{polygon}}$, must also agree on the extensions of the relations \textit{inside} and \textit{outside}. $\Box$

We are now in a position to prove the axiomatizability theorem for models of $T_{scene} \cup T_{fg}$.

Proof: By Lemma 4.3, all models of $T_{scene} \cup T_{fg}$ satisfy conditions 3.4.5. and 7 in the definition of $M_{fg}$.

By Lemma 4.10, all models of $T_{scene} \cup T_{fg}$ satisfy condition 6 in the definition of $M_{fg}$.

By Lemma 4.9, all models of $T_{scene} \cup T_{fg}$ satisfy conditions 8 and 9 in the definition of $M_{fg}$.

$\Box$

4.3.6 Independence of Axioms in $T_{fg}$

Theorem 4.6 shows that the axioms of $T_{fg}$ are sufficient to prove that there is a unique figure ground assignment for a surface. We can also show that they are necessary insofar as we can find counterexamples to the theorem if any axiom is falsified, that is, we can find a set of connected edges that either have ambiguous figure ground assignments, or no consistent figure ground assignment. This also demonstrates the independence of the axioms in $T_{fg}$, since a structure with these properties is either not a model of $T_{fg}$, or it cannot be a member of $M_{fg}$.

Theorem 4.8 The axioms in $T_{fg}$ is a minimal set necessary to ensure that every surface has a unique figure/ground assignment.

Proof: Consider the surface represented in Figure 4.27(a) where there exists a vertex such that three edges meet at that vertex. For any variable assignment $\sigma$, we have

$$M_1, \sigma \models \text{outside}(e_1, e_2) \land \text{outside}(e_2, e_3) \land \text{outside}(e_3, e_4) \land \text{outside}(e_4, e_1) \land \text{inside}(e_1, e_4)$$

so that $T_{scene} \cup T_{fg} - 4.2$ is inconsistent.

Consider the surface represented in Figure 4.27(b). There exist models $M_1, M_2$ of $T_{scene} \cup T_{fg} - 4.10$ with domain $\{e_1, e_2, e_3, e_4, e_5, e_6\} \subset M$ such that

1. $$(e_2, e_1) \in \text{inside}^{M_1}, (e_3, e_2) \in \text{inside}^{M_1}, (e_4, e_3) \in \text{inside}^{M_1}$$
$$ (e_5, e_4) \in \text{inside}^{M_1}, (e_6, e_5) \in \text{inside}^{M_1}$$

2. $$\langle e_2, e_3 \rangle \in \text{inside}^{M_2}, \langle e_3, e_2 \rangle \in \text{outside}^{M_2}, \langle e_4, e_3 \rangle \in \text{inside}^{M_2}$$
$$ (e_5, e_4) \in \text{inside}^{M_2}, (e_6, e_5) \in \text{inside}^{M_2}, (e_1, e_6) \in \text{inside}^{M_2}$$

$\textbf{In this section, the relations inside and outside have superscripts to distinguish their extensions in different structures.}$
Figure 4.27: Counterexamples for the axioms in $T_{fg}$. 
These anomalous figure/ground assignments are illustrated in Figure 4.27(c).

Consider the surface represented in Figure 4.27(d). There exist models $\mathcal{M}_1, \mathcal{M}_2, \mathcal{M}_3, \mathcal{M}_4$ of $T_{\text{scene}} \cup T_{fg} = 4.11$ and $T_{\text{scene}} \cup T_{fg} = 4.12$ with domain $\{e_1, e_2, e_3, e_4\} \subset M$ such that:

1. $\langle e_1, e_3 \rangle \in \text{inside}^{\mathcal{M}_1}, \langle e_2, e_4 \rangle \in \text{inside}^{\mathcal{M}_1}$
2. $\langle e_1, e_3 \rangle \notin \text{inside}^{\mathcal{M}_2}, \langle e_2, e_4 \rangle \notin \text{inside}^{\mathcal{M}_2}$
3. $\langle e_1, e_3 \rangle \in \text{inside}^{\mathcal{M}_3}, \langle e_2, e_4 \rangle \notin \text{inside}^{\mathcal{M}_3}$
4. $\langle e_1, e_3 \rangle \notin \text{inside}^{\mathcal{M}_4}, \langle e_2, e_4 \rangle \in \text{inside}^{\mathcal{M}_4}$

Consider the surface represented in Figure 4.27(b). There exist models $\mathcal{M}_1, \mathcal{M}_2, \mathcal{M}_3, \mathcal{M}_4$ of $T_{\text{scene}} \cup T_{fg} = 4.13$ with domain $\{e_1, e_2, e_3, e_4\} \subset M$ such that:

1. $\langle e_1, e_2 \rangle \in \text{inside}^{\mathcal{M}_1}, \langle e_2, e_3 \rangle \in \text{inside}^{\mathcal{M}_1}, \langle e_3, e_1 \rangle \in \text{inside}^{\mathcal{M}_1}$
   $\langle e_1, e_2 \rangle \in \text{outside}^{\mathcal{M}_1}, \langle e_2, e_3 \rangle \in \text{outside}^{\mathcal{M}_1}, \langle e_3, e_1 \rangle \in \text{outside}^{\mathcal{M}_1}$
2. $\langle e_1, e_2 \rangle \in \text{inside}^{\mathcal{M}_2}, \langle e_2, e_3 \rangle \in \text{inside}^{\mathcal{M}_2}, \langle e_3, e_1 \rangle \in \text{inside}^{\mathcal{M}_2}$
   $\langle e_1, e_2 \rangle \notin \text{outside}^{\mathcal{M}_2}, \langle e_2, e_3 \rangle \notin \text{outside}^{\mathcal{M}_2}, \langle e_3, e_1 \rangle \notin \text{outside}^{\mathcal{M}_2}$
3. $\langle e_1, e_2 \rangle \in \text{inside}^{\mathcal{M}_3}, \langle e_2, e_3 \rangle \in \text{inside}^{\mathcal{M}_3}, \langle e_3, e_1 \rangle \in \text{inside}^{\mathcal{M}_3}$
   $\langle e_1, e_2 \rangle \in \text{outside}^{\mathcal{M}_3}, \langle e_2, e_3 \rangle \notin \text{outside}^{\mathcal{M}_3}, \langle e_3, e_1 \rangle \in \text{outside}^{\mathcal{M}_3}$
4. $\langle e_1, e_2 \rangle \in \text{inside}^{\mathcal{M}_4}, \langle e_2, e_3 \rangle \in \text{inside}^{\mathcal{M}_4}, \langle e_3, e_1 \rangle \in \text{inside}^{\mathcal{M}_4}$
   $\langle e_1, e_2 \rangle \notin \text{outside}^{\mathcal{M}_4}, \langle e_2, e_3 \rangle \in \text{outside}^{\mathcal{M}_4}, \langle e_3, e_1 \rangle \notin \text{outside}^{\mathcal{M}_4}$

Lemma 3.3 illustrates the necessity of axiom 4.14; without it, there could exist surfaces such that $T_{fg}$ will have two models with different extensions of inside and outside, and there is no way to eliminate the ambiguity. Similarly, without axiom 4.5, it would be consistent for surfaces to have multiple disconnected sets of outer edges, and axiom 4.14 would be unable to force a unique extension of inside and outside. □

Thus we have a theory of shape which is minimal, yet also powerful enough to prove that all surfaces have a unique shape as defined by the assignment of figure and ground relations to the edges of the surface. We now proceed to show how this property is preserved by depiction.
Chapter 5

Depiction Axioms for Figure and Ground

In this chapter, we will characterize the class of structures which capture the relationship between figure/ground assignments and depicted surfaces. We will then present the axioms in the CardWorld module $T_\Delta$, and show that the models of $T_\Delta \cup T_{fg} \cup T_{kernel} \cup T_{scene}$ are equivalent to the class of structures which we introduced. In particular, we will be considering how the figure/ground relations (inside and outside) are preserved by depiction. In the preceding chapter, we proved that all surfaces have a unique figure ground assignment; the characterization theorem in this chapter will prove that depiction preserves this unique assignment.

The axioms of $T_\Delta$ are defined by their semantic properties, not by their syntactic form. This is an important feature, since several approaches ([Poole 87], [Reiter and Mackworth 89]) have represented depiction as a mapping from scene to image, or image to scene. This often gives the depiction axioms the syntactic form of Horn sentences in which all negative literals are scene literals and all positive literals are image literals. The depiction axioms in this chapter do not all have this implicit directionality. Rather, the central requirement for $T_\Delta$ will be to capture the following intuition:

Intuition 8 Depiction preserves the unique figure/ground assignment for edges.

With this intuition, we can do two things. Given a set of regions which we know to depict some set of surfaces, we can determine the figure/ground assignment for the depicted edges. Conversely, since depicted surfaces can be given a unique assignment of figure and ground, we can use this assignment to uniquely determine which regions depict which surfaces.

In terms of images, this property is closely related to the notion of ambiguity in an image. If we have already grouped the lines in an image by asserting that they depict edges in the same surface, then the only ambiguity over the depiction of surfaces is due to uncertainty in the figure ground relation of the depicted edges; for example, if we do not know whether $inside(e_1, e_2)$ or $outside(e_1, e_2)$, then we cannot assign surface depiction literals.
5.1 Extending the Language for Images

The nonlogical lexicon \( \mathcal{L}_\Delta \) which we will adopt for depicted surfaces will be

\[
\mathcal{L}_\Delta = \mathcal{L}_{\text{kernel}} \cup \mathcal{L}_{fg} \cup \{ \theta_i(a, l_1, l_2), \text{parallel}(l_1, l_2), \text{between}(l_1, l_2, r) \}
\]

- The ternary predicate \( \theta_i(a, l_1, l_2) \) denotes the relation whose intended interpretation is that the relative alignment between the lines \( l_1 \) and \( l_2 \) is \( a \). Intuitively, \( \theta_i(a, l_1, l_2) \) iff we can clockwise rotate the line containing the segment for \( l_1 \) by an angle of \( a \) radians into the line containing the segment for \( l_2 \), using the intersection point of their extensions as the pivot. If the intersection point is not part of either segment, then the alignment value is defined modulo \( 2\pi \) radians. If the intersection point is part of one of the segments, then the alignment value is defined modulo \( \pi \) radians.

- The binary predicate \( \text{parallel}(l_1, l_2) \) denotes the relation whose intended interpretation is that the lines \( l_1, l_2 \) are parallel in the image.

- The ternary predicate \( \text{between}(l_1, l_2, r) \) denotes the relation defined over two lines \( (l_1, l_2) \) and a region \( (r) \). The intended interpretation is that \( \text{between}(l, l', r) \) iff region \( r \) is between lines \( l \) and \( l' \) as we rotate clockwise from \( l \) to \( l' \) using the intersection point of their extensions as the pivot when determining the relative alignment of \( l \) and \( l' \).

Note that the only expansion of the language in this chapter is the introduction of these two new image predicates.

Example: In Figure 5.1(a) we have

\[
I \models \text{between}(l_1, l_6, r_1) \land \text{between}(l_2, l_1, r_1) \land \text{between}(l_1, l_7, r_1) \land \text{between}(l_1, l_8, r_1) \land \neg \text{between}(l_1, l_8, r_1)
\]
\[
\land \text{between}(l_4, l_3, r_2) \land \text{between}(l_4, l_2, r_2) \land \text{between}(l_5, l_4, r_2) \land \neg \text{between}(l_7, l_4, r_2)
\]

\[\square\]

The relations \( \theta_i \) and \( \text{parallel} \) are analogous to \( \theta \) and \( \text{parallel} \), respectively, defined in the previous chapter: the only difference is that \( \theta_i \) and \( \text{parallel} \) are image relations over lines, whereas \( \theta \) and \( \text{parallel} \) are scene relations over edges.

As with the polygon relations in \( \mathcal{L}_{\text{polygon}} \), we will formally specify the intended interpretations of \( \theta_i \) and \( \text{between} \) using Hilbert's axiomatization of geometry. We will define a new class of structures \( \mathcal{M}^{\text{angle}}_i \) which capture these intended interpretations.

Definition 5.1 Let \( \mathcal{H} \) be the model of Hilbert's geometry, and suppose \( \mathcal{M} \in \mathcal{L}_\Delta \). Then \( \mathcal{M} \in \mathcal{M}_i^{\text{angle}} \) iff

- If the lines containing the segments \( h(l_1) \) and \( h(l_2) \) intersect, then \( (l_1, l_2, r) \in \text{between} \iff (l_1, r) \in \text{in} \) and there exists a pixel \( q \) such that \( (q, r) \in \text{in} \) and \( h(q) \) is in the interior of the angle formed by the lines in \( \mathcal{H} \) containing the segments \( h(l_1) \) and \( h(l_2) \).

- If the lines containing the segments \( h(l_1) \) and \( h(l_2) \) are parallel, then \( (l_1, l_2, r) \in \text{between} \iff (l_1, r) \in \text{in} \) and there exists a pixel \( q \) such that \( (q, r) \in \text{in} \) and there exists a segment containing a point \( x_1 \) in \( h(l_1) \), a point \( x_2 \) in \( h(l_2) \) such that \( (x_1, h(q), x_2) \in \text{ordered} \).
• \((a, l_1, l_2) \in \theta_i \iff \langle h(l_1), h(l_2) \rangle \in \text{angle}\)

• \((l_1, l_2) \in \parallelM \iff \langle h(l_1), h(l_2) \rangle \in \parallelH\), i.e., lines are parallel in \(M\) iff the segments in \(H\) corresponding to the lines are also parallel in \(H\).

It is easy to see that the extensions of \(\theta_i\) and \(\parallel_i\) are isomorphic to the following class of graphs:

**Definition 5.2** The \(\theta_i\) graph for an image \(I\) is a complete graph with the following properties:

1. Each node of the graph corresponds to a line in the image \(I\).
2. The graph is weighted, that is, there is a unique real number associated with each directed arc in the graph.
3. If \((l_1, l_2)\) is a directed arc in the graph with weight \(\alpha\), and the intersection point of \(h(l_1)\) and \(h(l_2)\) is not contained in either of these lines, then the directed arc \((l_2, l_1)\) in the graph has the weight \(2\pi - \alpha\).
4. If \((l_1, l_2)\) is a directed arc in the graph with weight \(\alpha\), and the intersection point of \(h(l_1)\) and \(h(l_2)\) is contained in either of these lines, then the directed arc \((l_2, l_1)\) in the graph has the weight \(\pi - \alpha\).
5. If \((l_1, l_2) \in \parallel_i\), then the weight of the arcs \((l_1, l_2)\) and \((l_2, l_1)\) are both equal to 0.

We can also characterize the class of graphs which are isomorphic to the extension of \(\text{between}\). Rather than considering the structures for the complete extension of \(\text{between}\), we will consider the following restricted class of structures, in which one of the lines in the extension of the relation is also contained in the region:

**Definition 5.3** Given a structure \(M\) in \(L\), the \text{between} graph for a region \(r \in M\) is a directed graph \(G = (N, A)\) with the following properties:

1. \(N\) is the set of lines in \(M\);
2. \(G\) is the union of a set of \(K_{1,n}\) subgraphs, such that the root node \(l_1\) of each subgraph is a line in the region \(r\), and there is an arc \((l_1, l_2) \in A \iff \langle l_1, l_2, r \rangle \in \text{between}\)

Example: In Figure 5.1, we have an image (Figure 5.1(a)) and its associated \text{between} graphs for region \(r_1\) (Figure 5.1(b)) and region \(r_2\) (Figure 5.1(c)).

### 5.2 Motivating Examples for Depiction

In this section, we examine the intuitions underlying the depiction of edges by lines satisfying the alignment relation \(\theta_i\) and the relation \text{between}. In particular, we will show how the extension of \text{between} is intimately related to the extension of the figure/ground assignments (extension of \text{inside} and \text{outside}) for surfaces depicted by the regions which are between lines depicting edges in those surfaces.
Figure 5.1: Images and their between graphs.
Figure 5.2: Depiction of alignment and indeterminate edges.
5.2.1 Depiction of Alignment

The first property of depicted edges that we notice is that the relative alignment of depicted indeterminate edges need not be fixed in all images – depending on which part of the edge is depicted, the angle between the depicting lines is different.

Example: Let $I_a, I_b, I_c$ be the images in Figure 5.2(a),(b),(c), respectively. Suppose that the surface whose edges are depicted is the one in Figure 5.2(e). There exists a structure $\mathcal{M} \in \mathcal{L}_\Delta$ such that $\{e_1, e_2, s_1\} \subset M$, and $\mathcal{M} \models I$, and

$$\langle e_1, s_1 \rangle, \langle e_2, s_1 \rangle \in \text{part}$$

$$\langle e_1, e_2 \rangle \in \text{indet}$$

$$\langle \frac{\pi}{4}, e_1, e_2 \rangle, \langle \frac{3\pi}{4}, e_2, e_1 \rangle \in \theta$$

$$\langle r_1, s_1 \rangle, \langle l_1, e_1 \rangle, \langle l_2, e_2 \rangle \in \Delta$$

All of the images in Figure 5.2 are consistent with this assignment of depiction literals. In (a) we have

$$I_a \models \theta_i(\frac{\pi}{4}, l_1, l_2) \land \theta_i(\frac{3\pi}{4}, l_2, l_1)$$

In (b) we have

$$I_b \models \theta_i(\frac{\pi}{4}, l_1, l_2) \land \theta_i(\frac{7\pi}{4}, l_2, l_1)$$

In (c) we have

$$I_c \models \theta_i(\frac{5\pi}{4}, l_1, l_2) \land \theta_i(\frac{3\pi}{4}, l_2, l_1)$$

$\square$

Example: Let $I$ be the image in Figure 5.2(d). There exists a structure $\mathcal{M} \in \mathcal{L}_\Delta$ such that $\{e_1, e_2, s_1\} \subset M$, and $\mathcal{M} \models I$, and

$$\langle e_1, s_1 \rangle, \langle e_2, s_1 \rangle \in \text{part}$$

$$\langle e_1, e_2 \rangle \in \text{indet}$$

$$\langle \frac{\pi}{4}, e_1, e_2 \rangle, \langle \frac{3\pi}{4}, e_2, e_1 \rangle \in \theta$$

$$\langle r_1, s_1 \rangle, \langle l_1, e_1 \rangle, \langle l_2, e_2 \rangle, \langle l_3, e_2 \rangle \in \Delta$$

Notice that we have

$$I \models \theta_i(\frac{\pi}{4}, l_1, l_2) \land \theta_i(\frac{3\pi}{4}, l_2, l_1) \land \theta_i(\frac{5\pi}{4}, l_1, l_3) \land \theta_i(\frac{3\pi}{4}, l_3, l_1)$$

In this case, we have two different lines depicting the same edge, yet the alignment of each line is different. $\square$
Figure 5.3: Collinear lines and the depiction of a single edge.
The second property of relative alignment for depicted edges can be seen when we consider a partially occluded edge. In this case, there are two distinct lines that depict the same edge: further, the relative alignment of these two lines is always $\pi$ i.e., lines depicting the same edge are always collinear. For example, in Figure 5.3(a), there exists a structure in which the lines $l_1$ and $l_2$ depict the same edge. Similarly, in Figure 5.3(b), there exists a structure in which the three lines $l_1, l_2$, and $l_3$ all depict the same edge.

5.2.2 Depiction and the between Relation

To motivate the relationship of between to the figure/ground assignment of a depicted surface, we will first consider some example images.

Example: Consider the image in Figure 5.4(a), in which

$I \models between(l_2, l_1, r_2) \land between(l_1, l_2, r_3) \land \neg between(l_2, l_1, r_4) \land \neg between(l_1, l_2, r_1)$

Suppose that the surface $s$ depicted in the images is the surface in Figure 5.5(a), so that there is a structure $\mathcal{M}_1$ in $\mathcal{L}_{\Delta}$ such that

$\langle e_6, e_5 \rangle, \langle e_3, e_2 \rangle \in \text{inside}$

$\langle e_5, e_9 \rangle, \langle e_2, e_3 \rangle \notin \text{inside}$

$\langle e_5, e_9 \rangle, \langle e_2, e_3 \rangle \notin \text{indet}$

Then we should have

$\langle l_1, e_8 \rangle, \langle l_2, e_5 \rangle \in \Delta$

iff

$\langle r_2, s \rangle, \langle r_1, s \rangle \in \Delta$

$\langle r_3, s \rangle, \langle r_4, s \rangle \notin \Delta$

This figure/ground assignment for depicted surfaces is pictorially presented in Figure 5.6(a), where the arrows indicate the region which depicts the surface containing the edge depicted by the line.

There is also a structure $\mathcal{M}_2$ in $\mathcal{L}_{\Delta}$ such that

$\langle l_1, e_2 \rangle, \langle l_2, e_3 \rangle \in \Delta$

iff

$\langle r_4, s \rangle, \langle r_3, s \rangle \in \Delta$

$\langle r_2, s \rangle, \langle r_1, s \rangle \notin \Delta$

This figure/ground assignment for depicted surfaces is pictorially presented in Figure 5.6(b). $\square$
Figure 5.4: Examples for depiction axioms.
Figure 5.5: Surfaces which are depicted in the images of Figure 5.4.
Figure 5.6: Figure/ground assignments for depicted surfaces.
Figure 5.7: Figure/ground assignments for depicted surfaces (cont.).
Example: Consider the image in Figure 5.4(c), in which

\[ I \models \text{between}(l_1, l_2, r_3) \land \text{between}(l_2, l_1, r_4) \land -\text{between}(l_1, l_2, r_1) \land -\text{between}(l_2, l_1, r_2) \]
\[ \land \theta\left(\frac{\pi}{4}, l_1, l_2\right) \]

Suppose that the surface \( s \) depicted in the images is the surface in Figure 5.5(b), so that there is a structure \( \mathcal{M}_1 \) in \( \mathcal{L}_\Delta \) such that

\[ \langle e_1, e_2 \rangle \in \text{inside}, \langle e_2, e_1 \rangle \notin \text{inside} \]
\[ \langle e_1, e_2 \rangle \in \text{indet} \]
\[ \langle e_1, e_2 \rangle \in \text{convex} \]

Then we should have

\[ \langle l_1, e_2 \rangle, \langle l_2, e_1 \rangle \in \Delta \]

iff

\[ \langle r_2, s \rangle, \langle r_1, s \rangle \in \Delta \]
\[ \langle r_3, s \rangle, \langle r_4, s \rangle \notin \Delta \]

This figure/ground assignment for depicted surfaces is pictorially presented in Figure 5.6(c).

Example: Consider the image in Figure 5.4(d), in which

\[ I \models \text{between}(l_1, l_2, r_3) \land \text{between}(l_2, l_1, r_4) \land -\text{between}(l_1, l_2, r_1) \land -\text{between}(l_2, l_1, r_2) \]
\[ \land \theta\left(\frac{5\pi}{4}, l_1, l_2\right) \]

Suppose that the surface \( s \) depicted in the images is the surface in Figure 5.5(b), so that there is a structure \( \mathcal{M}_1 \) in \( \mathcal{L}_\Delta \) such that

\[ \langle e_1, e_2 \rangle \in \text{inside}, \langle e_2, e_1 \rangle \notin \text{inside} \]
\[ \langle e_1, e_2 \rangle \in \text{indet} \]
\[ \langle e_1, e_2 \rangle \in \text{convex} \]

Then we should have

\[ \langle l_1, e_2 \rangle, \langle l_2, e_1 \rangle \in \Delta \]

iff

\[ \langle r_2, s \rangle, \langle r_1, s \rangle \in \Delta \]
\[ \langle r_3, s \rangle, \langle r_4, s \rangle \notin \Delta \]

This figure/ground assignment for depicted surfaces is pictorially presented in Figure 5.7(e).
Example: Consider the image in Figure 5.4(c), in which

\[ I \models between(l_1, l_2, r_3) \land between(l_2, l_1, r_4) \land \neg between(l_1, l_1, r_1) \land \neg between(l_2, l_2, r_2) \]
\[ \land \theta(l, l_1, l_2) \]

Suppose that the surface \( s \) depicted in the images is the surface in Figure 5.5(c), so that there is a structure \( M_1 \) in \( L_\Delta \) such that
\[
\langle e_1, e_2 \rangle \notin \text{inside}, \langle e_2, e_1 \rangle \notin \text{inside}
\]
\[
\langle e_1, e_2 \rangle \in \text{indet}
\]
\[
\langle e_1, e_2 \rangle \notin \text{convex}
\]

Then we should have
\[
\langle l_1, e_2 \rangle, \langle l_2, e_1 \rangle \in \Delta
\]
iff
\[
\langle r_4, s \rangle, \langle r_1, s \rangle \in \Delta
\]
\[
\langle r_2, s \rangle, \langle r_3, s \rangle \notin \Delta
\]

This figure/ground assignment for depicted surfaces is pictorially presented in Figure 5.7(d).

Example: Consider the image in Figure 5.4(d), in which

\[ I \models between(l_1, l_2, r_3) \land between(l_2, l_1, r_4) \land \neg between(l_1, l_2, r_1) \land \neg between(l_2, l_1, r_2) \]
\[ \land \theta(l, l_1, l_2) \]

Suppose that the surface \( s \) depicted in the images is the surface in Figure 5.5(c), so that there is a structure \( M_1 \) in \( L_\Delta \) such that
\[
\langle e_1, e_2 \rangle, \langle e_2, e_1 \rangle \notin \text{inside}
\]
\[
\langle e_1, e_2 \rangle \in \text{indet}
\]
\[
\langle e_1, e_2 \rangle \notin \text{convex}
\]

Then we should have
\[
\langle l_1, e_2 \rangle, \langle l_2, e_1 \rangle \in \Delta
\]
iff
\[
\langle r_4, s \rangle, \langle r_1, s \rangle \in \Delta
\]
\[
\langle r_2, s \rangle, \langle r_3, s \rangle \notin \Delta
\]

This figure/ground assignment for depicted surfaces is pictorially presented in Figure 5.7(f).

We will next introduce various classes of structures to formalize these intuitive figure/ground assignments for depicted surfaces.
5.3 Structures for Depicted Surfaces

We will now present various classes of structures which will be used to characterize the models of the depiction axioms for figure and ground. In particular, we present two classes of structures related to the two new image relations in \( L_\Delta \). We first will define the structures that capture the relationship between \( \theta \) (scene relation) and \( \theta_i \) (image relation). We then define the structures that capture the intuitions about the relationship among the figure/ground assignments of depicted surfaces and the image relation between.

5.3.1 Relative Alignment and Depicted Edges

The first class of structures we consider will characterize the relationship between the scene relation \( \theta(a, e_1, e_2) \) and the image relation \( \theta_i(a, l_1, l_2) \) in the case where the lines \( l_1, l_2 \) depict the edges \( e_1, e_2 \).

**Definition 5.4** Given a structure \( M \) in \( L_\Delta \), the depicted \( \theta_i \) subgraph for a surface \( s \) \( \in M \) is the subgraph of the graph isomorphic to the extension of \( \theta_i \) containing all lines in \( M \) such that for each line \( l \) there exists an edge \( e \) \( \in M \) such that

\[
(174 E A \text{ in other words}, \text{given a structure } M \in L_\Delta, \text{ the depicted } \theta_i \text{ subgraph for a surface contains only lines which depict edges in the surface in } M.)

**Definition 5.5** Given a structure \( M \) in \( L_\Delta \), the depicted alignment graph is a graph satisfying the following properties:

1. The graph is the union of depicted \( \theta_i \) subgraphs for each depicted surface in \( M \).
2. If \( (l_1, e_1) \in \Delta \) and \( (l_2, e_2) \in \Delta \) and \( (e_1, e_2) \in \text{parallel} \), then the weight of the directed arc \( (l_1, l_2) \) in the depicted \( \theta_i \) subgraph is 0.
3. If \( (l_1, e) \in \Delta \) and \( (l_2, e) \in \Delta \) for two distinct lines \( l_1, l_2 \) \( \in M \), then the weight of the directed arc \( (l_1, l_2) \) in the depicted \( \theta_i \) subgraph is \( \pi \).
4. If \( (l_1, e_1) \in \Delta \) and \( (l_2, e_2) \in \Delta \) and \( (e_1, e_2) \in \text{indet} \) and \( (\alpha, e_1, e_2) \in \theta \) then \( \alpha \) is the weight of the directed arc \( (l_1, l_2) \) in the depicted \( \theta_i \) subgraph.
5. If \( (l_1, e_1) \in \Delta \) and \( (l_2, e_2) \in \Delta \) and \( (e_1, e_2) \in \text{indet} \) and \( (\alpha, e_1, e_2) \in \theta \) then either \( \alpha \) or \( \alpha + \pi \) is the weight of the directed arc \( (l_1, l_2) \) in the depicted \( \theta_i \) subgraph.

For example, consider the scenes in Figure 5.4(c) and (d), supposing that the surface in Figure 5.4(c) is the one which is being occluded. The depicted alignment graph for the surface in Figure 5.4(c) is shown in Figure 5.8(a). The depicted alignment graph for the surface in Figure 5.4(d) is shown in Figure 5.8(b).

In Figure 5.3(a), there exists a structure in which the lines \( l_1 \) and \( l_2 \) depict the same edge. The depicted alignment graph for this structure is shown in Figure 5.9(a). In Figure 5.3(b), there exists a structure in which the lines \( l_1, l_2, \) and \( l_3 \) depict the same edge. The depicted alignment graph for this structure is shown in Figure 5.9(b).
Figure 5.8: Depicted alignment graphs for images in Figure 5.4.

Figure 5.9: Depicted alignment graphs for image in Figure 5.3(a).
Figure 5.10: Depicting between graphs for image in Figure 5.1.

5.3.2 Depicting Figure and Ground

As we saw with the motivating examples, the depiction of inside and outside is tied to the extension of between for lines depicting edges in the same surface. Under certain conditions, the figure/ground assignment for edges in a surface is preserved by the between relation over the lines which depict the edges. Similarly, the figure/ground assignments for a set of edges in a surface can be used to determine which region depicts that surface, based on the extension of between for the lines depicting the edges. The following classes of structures formalize this intuition.

In the first class of structures, we are extracting the subgraph of a between graph which contains only the lines which depict edges in the same surface.

**Definition 5.6** Given a structure $M$ in $L_\Delta$, the depicting between subgraph for a region $r \in M$ is the subgraph of the between graph containing all lines in $M$ such that for each line $l$ there exists an edge $e \in M$ such that $(l,e) \in \Delta$, and there exists a surface $s$ such that for each of these edges, we have $(e,s) \in \text{part}$.

**Example:** Consider the image in Figure 5.1(a). There is a structure $M$ in $L_\Delta$ such that

$\langle r_1, s_1 \rangle, \langle r_2, s_2 \rangle \in \Delta$

The depicting between graph for $r_1$ is shown in Figure 5.10(a), and the depicting between graph for $r_2$ is shown in Figure 5.10(b). □

In the second class of structures, we are extracting the subgraph of a figure/ground graph which contains only the edges which are depicted in the image. Note that multiple edges of the surface must be depicted for this subgraph to exist.

**Definition 5.7** Given a structure $M$ in $L_\Delta$, the depicted inside subgraph for a surface $s \in M$ is the subgraph of the inside graph containing all lines in $M$ such that for each edge $e$ there exists a line $l \in M$ such that $(l,e) \in \Delta$ and $(l,r) \in \text{in}$. 
These two classes of subgraphs can now be used to define the classes of structures which characterize the depiction of figure and ground relations within an image. The definition specifies the conditions under which depiction of a surface either preserves or complements the figure/ground assignment for the depicted edges in the surface.

**Definition 5.8** Given a structure M in LA, the depicted figure/ground graph for a depicted surface s ∈ M is a graph G with the following properties: 1

1. The graph G is the union of the depicted inside subgraph G₁ for the surface and the depicting between subgraph G₂ for a region r ∈ M, such that each edge that is an element of the depicted inside subgraph is depicted by a line that is an element of the depicting between subgraph.

2. If (r, s) ∈ Δ and (e₁, e₂) ∉ indet, then

   \[(e₁, e₂) ∈ G₁ \text{ iff } (l₁, l₂) ∈ G₂\]

3. If (r, s) ∉ Δ and (e₁, e₂) ∉ indet, then

   \[(e₁, e₂) ∈ G₁ \text{ iff } (l₁, l₂) ∉ G₂\]

4. If (r, s) ∈ Δ and (e₁, e₂) ∈ indet, and (a, l₁, l₂) ∈ θᵢ such that a < π, then

   \[(e₁, e₂) ∈ G₁ \text{ iff } (l₁, l₂) ∈ G₂\]

5. If (r, s) ∉ Δ and (e₁, e₂) ∈ indet, and (a, l₁, l₂) ∈ θᵢ such that a < π, then

   \[(e₁, e₂) ∈ G₁ \text{ iff } (l₁, l₂) ∉ G₂\]

6. If (r, s) ∈ Δ and (e₁, e₂) ∈ indet, and (a, l₁, l₂) ∈ θᵢ such that a > π, then

   \[(e₁, e₂) ∈ G₁ \text{ iff } (l₁, l₂) ∉ G₂\]

7. If (r, s) ∉ Δ and (e₁, e₂) ∈ indet, and (a, l₁, l₂) ∈ θᵢ such that a > π, then

   \[(e₁, e₂) ∈ G₁ \text{ iff } (l₁, l₂) ∈ G₂\]

In Figures 5.11 - 5.16, we illustrate each of the cases in the definition of depicted figure/ground graphs. In each graph, the edges form a substructure of the depicted inside graph, and the lines form a substructure of the depicting between graph. An arc between two edges e₁, e₂ indicates \((e₁, e₂) ∈ \text{inside}\) in the surface s, while absence of an arc indicates \((e₁, e₂) ∉ \text{inside}\) in the surface s. Similarly, an arc between two lines l₁, l₂ indicates \((l₁, l₂, r) ∈ \text{between}\), while absence of an arc indicates \((l₁, l₂, r) ∉ \text{between}\). Together, each graph in Figures 5.11 - 5.16 is a substructure of the depicted figure/ground graph in some structure. Of course, the entire figure/ground graph would combine such graphs for each pair of depicted edges in a surface.

In particular, the graphs in Figures 5.11 - 5.16 are the depicted figure/ground graphs in structures corresponding to the images in Figure 5.4 and their intuitive figure/ground assignments in Figures 5.6 and 5.6. The depicted figure/ground graphs for Figure 5.6(a) are found in Figure 5.11(a) and Figure

---

1In all of the following conditions, we are assuming \((l₁, e₁), (l₂, e₂) ∈ Δ\).
Figure 5.11: Subgraphs of depicted figure/ground graphs corresponding to property (2) in the definition of depicted figure/ground graphs. The objects in (a) are from Figure 5.6(a), and the objects in (b) are from Figure 5.6(b).

5.12(a). The depicted figure/ground graphs for Figure 5.6(b) are found in Figure 5.11(b) and Figure 5.12(b). The depicted figure/ground graphs for Figure 5.6(c) are found in Figure 5.13(a) and Figure 5.14(a). The depicted figure/ground graphs for Figure 5.7(d) are found in Figure 5.13(b) and Figure 5.14(b). The depicted figure/ground graphs for Figure 5.7(e) are found in Figure 5.15(a) and Figure 5.16(a). The depicted figure/ground graphs for Figure 5.7(f) are found in Figure 5.15(b) and Figure 5.16(b).

We can notice that the depicted figure/ground graph can only exist if the surface has multiple depicted edges. However, there are cases where a surface is so occluded that only one depicted edge of the surface exists. In this case, we need to define an additional class of structures:

Definition 5.9 A solitary depicting graph is a subgraph of the depicting between graph such that all regions depict the same surface iff the subgraph is a directed acyclic graph that is transitive and which has only one connected component.

Definition 5.10 Let $\mathcal{M}_i^\Delta$ be a set of structures with the following properties: For each structure $\mathcal{M} \in \mathcal{M}_i^\Delta$,

1. $\mathcal{M}$ is an extension of a structure in $\mathcal{M}_i^{kernel}$;
2. $\mathcal{M}$ is an extension of a structure in $\mathcal{M}_i^{fg}$;
3. $\mathcal{M}$ is an extension of a structure in $\mathcal{M}_i^{angle}$;
4. $\mathcal{M}$ contains a unique depicted figure/ground graph for each depicted surface in $\mathcal{M}$ that has multiple depicted edges, and a unique depicted alignment graph for each depicted surface in $\mathcal{M}$ that has multiple depicted edges.
Figure 5.12: Subgraphs of depicted figure/ground graphs corresponding to property (3) in the definition of depicted figure/ground graphs. The objects in (a) are from Figure 5.6(a), and the objects in (b) are from Figure 5.6(b).

Figure 5.13: Subgraphs of depicted figure/ground graphs corresponding to property (4) in the definition of depicted figure/ground graphs. The objects in (a) are from Figure 5.6(c), and the objects in (b) are from Figure 5.7(d).
Figure 5.14: Subgraphs of depicted figure/ground graphs corresponding to property (5) in the definition of depicted figure/ground graphs. The objects in (a) are from Figure 5.6(c), and the objects in (b) are from Figure 5.7(d).

Figure 5.15: Subgraphs of depicted figure/ground graphs corresponding to property (6) in the definition of depicted figure/ground graphs. The objects in (a) are from Figure 5.7(e), and the objects in (b) are from Figure 5.7(f).
Figure 5.16: Subgraphs of depicted figure/ground graphs corresponding to property (7) in the definition of depicted figure/ground graphs. The objects in (a) are from Figure 5.7(e), and the objects in (b) are from Figure 5.7(f).

Figure 5.17: Solitary depicting graphs for the images in Figure 5.3.
5. $M$ contains a unique solitary depicting graph for each depicted surface in $M$ that has a unique depicted edge.

By the first condition, every structure contains an image substructure and a scene substructure, and a depiction relation among image and scene elements. By the second condition, every structure is composed of surfaces with figure-ground assignments.

5.3.3 Characterization Theorems

We first establish that the class of structures in $M^\Delta_i$ exists and is nonempty.

**Theorem 5.1** Existence Theorem for $M^\Delta_i$

The class of structures $M^\Delta_i$ exists and is nonempty.

**Proof:** Since structures in $M^\Delta_i$ are composed of depicted figure/ground graphs and solitary depicting graphs for the surfaces in the domain, it suffices to show that these classes of graphs exist.

The depicted inside subgraph exists, since it is a subgraph of the inside graph for the surface, which exists by the Existence Theorem for $M^\Delta_i$.

The depicting between subgraph exists, since it is a subgraph of the between graph, which exists since it is the union of $K_{1,n}$ graphs and the extension of between is nonempty in models of Hilbert’s geometry.

The depicted figure/ground graphs are the union of these subgraphs, so the class of depicted figure/ground graphs is nonempty. Note that the conditions in the definition of depicted figure/ground graphs are simply constraining which subgraphs in each class are being combined.

The solitary depicting graphs are isomorphic to arbitrary directed chordal graphs, and hence they also exist. □

The next two theorems show that the figure/ground assignments for surfaces is preserved in depicted surfaces.

**Theorem 5.2** Structures in $M^\Delta_i$ which agree on the extension of $\Delta$ for edges in the same surface, also agree on the extension of inside and outside for these edges.

**Proof:** Suppose that there exist two structures $M_1, M_2$ in $M^\Delta_i$ which agree on the extension of $\Delta$ for edges in the same surface, but which disagree on the extension of inside and outside for these edges. Then $M_1$ and $M_2$ must have isomorphic depicted figure/ground graphs for some surface but nonisomorphic figure/ground graphs for the surface. However, the depicted figure/ground graphs are the subgraphs of the inside and outside graphs containing the depicted edges in the surface, so the inside and outside graphs must be isomorphic. □

**Theorem 5.3** Structures in $M^\Delta_i$ which agree on the extension of $\Delta$ for edges in the same surface and which agree on the extension of inside and outside, also agree on the extension of $\Delta$ for regions and surfaces.

**Proof:** Suppose that there exist two structures $M_1, M_2$ in $M^\Delta_i$ which agree on the extension of $\Delta$ for edges in the same surface, and which agree on the extension of inside and outside for these
edges, but which disagree on the set of depicted surfaces. Then \( M_1 \) and \( M_2 \) must have isomorphic depicted figure/ground graphs for all surfaces and isomorphic figure/ground graphs for the surfaces, but disagree on the set of depicted surfaces. However, each surface has a unique figure/ground graph and each depicted surface has a unique depicted figure/ground graph, so that \( M_1 \) and \( M_2 \) must agree on the set of depicted surfaces. □

### 5.4 The Axioms of \( T_\Delta \)

Let \( T_\Delta \) be the set of axioms in Figure 5.18.

First we have the depiction axioms for alignment relations. Axiom 5.1 of \( T_\Delta \) defines depiction for edges that are not indeterminate. The only difficulty here is given by indeterminate edges. As we observed earlier, the angle between the depicting lines is different, depending on which part of the indeterminate edge is depicted.

Thus we need a disjunctive depiction axiom for the relative alignment of indeterminate edges (axiom 5.2 of \( T_\Delta \)). This is an important distinction that has not been captured in previous work ([Grimson 90], [Jacobs 88]).

The depiction axioms allow us to determine the extension of the figure ground predicate \textit{inside}, given the extension of \textit{between}. We can illustrate axioms 5.5 and 5.6 using the images in Figure 5.4.

---

**Example:** Consider the image \( I \) in Figure 5.4(a) where

\[
I \models \text{between}(l_1, l_2, r_3) \land \neg \text{between}(l_1, l_2, r_1) \land \text{between}(l_2, l_1, r_2) \land \neg \text{between}(l_2, l_1, r_1)
\]

There exists a structure \( M \in M^3_i \) with \( \{e_1, e_2, s_1\} \subseteq M \) such that \( M \models I \) and

\[
\langle l_1, e_1 \rangle, \langle l_2, e_2 \rangle \in \Delta
\]

\[
\langle e_1, s_1 \rangle, \langle e_2, s_1 \rangle \in \text{part}
\]

\[
\langle r_1, s_1 \rangle, \langle r_2, s_1 \rangle \in \Delta
\]

\[
\langle e_1, e_2 \rangle \notin \text{indet}
\]

By axiom 5.5 of \( T_\Delta \) we can recover the figure ground assignment for \( e_1, e_2 \), since they are edges in a depicted surface:

\[
\langle e_2, e_1 \rangle \in \text{inside}, \langle e_2, e_1 \rangle \notin \text{outside}
\]

□

**Example:** Consider the image \( I \) in Figure 5.4(b) where

\[
I \models \text{between}(l_1, l_2, r_3) \land \neg \text{between}(l_1, l_2, r_1) \land \text{between}(l_2, l_1, r_2) \land \neg \text{between}(l_2, l_1, r_4)
\]
If two edges are not indeterminate, then the alignment of the edges is preserved by depiction of the edges.

\[(\forall e, e', s, l, l', \alpha) \quad \Delta(l, e) \land \Delta(l', e') \land \text{part}(e, s) \land \text{part}(e', s) \land \neg \text{indet}(e, e') \supset \]
\[\quad (\theta(\alpha, e, e') \equiv \theta_1(\alpha, l, l')) \quad (5.1)\]

If two edges are indeterminate, then the alignment of the lines depicting the edges is either equal to the alignment of the edges or the alignment of the edges incremented by \(\pi\).

\[(\forall e, e', s, l, l', \alpha) \quad \Delta(l, e) \land \Delta(l', e') \land \text{part}(e, s) \land \text{part}(e', s) \land \text{indet}(e, e') \supset \]
\[\quad (\theta(\alpha, e, e') \equiv (\theta_1(\alpha, l, l') \lor \theta_1(\pi + \alpha, l, l'))) \quad (5.2)\]

Parallel edges are depicted by parallel lines.

\[(\forall e, e', s, l, l') \quad \Delta(l, e) \land \Delta(l', e') \land \text{part}(e, s) \land \text{part}(e', s) \supset \]
\[\quad (\text{parallel}(e, e') \equiv \text{parallel}(l, l')) \quad (5.3)\]

Two lines depicting the same edge are collinear.

\[(\forall l, l, e, e', s) \quad \Delta(l, e) \land \Delta(l, e') \land l \neq l_2 \supset \theta_1(\pi, l, l_2) \quad (5.4)\]

If two distinct edges \(e_1, e_2\) are not indeterminate, then the surface is on the "\(e_2\) side" of \(e_1\) iff the region depicting the surface is between the lines depicting the edges. Equivalently, if two distinct edges \(e_1, e_2\) are not indeterminate, and some region depicts the surface containing these edges, then the surface is on the "\(e_2\) side" of \(e_1\) iff the region depicting the surface is between the lines depicting the edges.

\[(\forall e, e', s, l, l', r) \quad \text{part}(e, s) \land \text{part}(e', s) \land \Delta(l, e) \land \Delta(l', e') \land \text{in}(l, r) \land \neg \text{indet}(e, e') \land e \neq e' \supset \]
\[\quad \text{inside}(e, e') \equiv (\Delta(r, s) \equiv \text{between}(l, l', r)) \quad (5.5)\]

If two distinct edges \(e_1, e_2\) are indeterminate, then the surface is on the convex "\(e_2\) side" of \(e_1\) iff the region depicting the surface is not between the lines depicting the edges.

\[(\forall e, e', s, l, l', r) \quad \text{part}(e, s) \land \text{part}(e', s) \land \text{indet}(e, e') \land \Delta(l, e) \land \Delta(l', e') \land \theta_1(a, l, l') \land \alpha < \pi \land \text{in}(l, r) \land e \neq e' \supset \]
\[\quad (\text{convex}(e, e') \equiv (\Delta(r, s) \equiv \text{between}(l, l', r)) \quad (5.6)\]

If two distinct edges \(e_1, e_2\) are indeterminate, then the surface is on the convex "\(e_2\) side" of \(e_1\) iff the region depicting the surface is not between the lines depicting the edges.

\[(\forall e, e', s, l, l', r) \quad \text{part}(e, s) \land \text{part}(e', s) \land \text{indet}(e, e') \land \Delta(l, e) \land \Delta(l', e') \land \theta_1(a, l, l') \land \alpha > \pi \land \text{in}(l, r) \land e \neq e' \supset \]
\[\quad (\text{convex}(e, e') \equiv (\Delta(r, s) \equiv \neg \text{between}(l, l', r)) \quad (5.7)\]

If two lines depict the same edge in some surface, then the regions depicting the surface are on a unique side of the lines depicting the edge.

\[(\forall l, l', r, r', e, s) \quad \text{in}(l, r) \land \text{in}(l', r') \land \Delta(l, e) \land \Delta(l', e) \land \text{part}(e, s) \supset \]
\[\quad [(\Delta(r, s) \equiv \Delta(r', s)) \equiv (\text{between}(l, l', r) \equiv \neg \text{between}(l', l, r'))] \quad (5.8)\]
There exists a structure \( M \in \mathcal{M}_i^\Delta \) with \( \{e_1, e_2, s_1\} \subset M \) such that \( M \models I \) and

\[
\langle l_1, e_1 \rangle, \langle l_2, e_2 \rangle \in \Delta
\]

\[
\langle e_1, s_1 \rangle, \langle e_2, s_1 \rangle \in \text{part}
\]

\[
\langle r_3, s_1 \rangle, \langle r_4, s_1 \rangle \in \Delta
\]

\[
\langle e_1, e_2 \rangle \notin \text{indet}
\]

By axiom 5.5 of \( T_\Delta \) we can recover the figure ground assignment for \( e_1, e_2 \), since they are edges in a depicted surface:

\[
\langle e_2, e_1 \rangle \in \text{outside}, \langle e_2, e_1 \rangle \notin \text{inside}
\]

\[\square\]

Example: For the image \( I \) in Figure 5.4(c) we will use axiom (4) of \( T_\Delta \).

\[
I \models \text{between}(l_1, l_2, r_3) \land \neg\text{between}(l_1, l_2, r_1) \land \text{between}(l_2, l_1, r_2) \land \neg\text{between}(l_2, l_1, r_4) \land \theta_i(\frac{5\pi}{4}, l_1, l_2)
\]

There exists a structure \( M \in \mathcal{M}_i^\Delta \) with \( \{e_1, e_2, s_1\} \subset M \) such that \( M \models I \) and

\[
\langle l_1, e_1 \rangle, \langle l_2, e_2 \rangle \in \Delta
\]

\[
\langle e_1, s_1 \rangle, \langle e_2, s_1 \rangle \in \text{part}
\]

\[
\langle r_1, s_1 \rangle, \langle r_2, s_1 \rangle \in \Delta
\]

\[
\langle e_2, e_1 \rangle \in \text{indet}
\]

By axiom 5.6 of \( T_\Delta \) we can recover the figure ground assignment for \( e_1, e_2 \), since they are edges in a depicted surface:

\[
\langle e_1, e_2 \rangle \in \text{outside}, \langle e_2, e_1 \rangle \in \text{inside}
\]

\[\square\]

Example: Consider the image \( I \) in Figure 5.4(c) where

\[
I \models \text{between}(l_1, l_2, r_3) \land \neg\text{between}(l_1, l_2, r_1) \land \text{between}(l_2, l_1, r_2) \land \neg\text{between}(l_2, l_1, r_4) \land \theta_i(\frac{5\pi}{4}, l_1, l_2)
\]

There exists a structure \( M \in \mathcal{M}_i^\Delta \) with \( \{e_1, e_2, s_1\} \subset M \) such that \( M \models I \) and

\[
\langle l_1, e_1 \rangle, \langle l_2, e_2 \rangle \in \Delta
\]

\[
\langle e_1, s_1 \rangle, \langle e_2, s_1 \rangle \in \text{part}
\]

\[
\langle r_1, s_1 \rangle, \langle r_2, s_1 \rangle \in \Delta
\]

\[
\langle e_2, e_1 \rangle \in \text{indet}
\]
By axiom 5.6 of $T_\Delta$ we can recover the figure ground assignment for $e_1, e_2$, since they are edges in a depicted surface:

$$\langle e_1, e_2 \rangle \in \text{outside}, \langle e_2, e_1 \rangle \in \text{inside}$$

□

We must also make a special case for edges that are depicted by multiple lines, which are not covered by the axioms in Figure 5.18 that require the lines to depict distinct edges. Intuitively we want the correct regions to depict a surface; the other depiction axioms formalize this intuition by using the unique figure ground assignment guaranteed by the characterization theorem of the previous chapter. For multiply depicted edges, this intuition is formalized by axiom 5.8 of $T_\Delta$.

Example: Axiom 5.8 of $T_\Delta$ can be illustrated by the image $I$ in Figure 5.3(a). Suppose

$$I \models \text{in}(l_1, r_1) \land \text{in}(l_1, r_3) \land \text{in}(l_2, r_2) \land \text{in}(l_2, r_4) \land \theta_4(l_1, l_2)$$

$$\land \text{between}(l_1, l_2, r_3) \land -\text{between}(l_1, l_2, r_1) \land \text{between}(l_2, l_1, r_2) \land -\text{between}(l_2, l_1, r_4)$$

There exists a structure $\mathcal{M} \in \mathcal{M}_D^\Delta$ with $\{e_1, s_1\} \subseteq M$ such that $\mathcal{M} \models I$ and

$$(l_1, e_1), (l_1, e_1) \in \Delta$$

$$\langle e_1, s_1 \rangle \in \text{part}$$

By axiom 5.8, there are two models $\mathcal{M}_1, \mathcal{M}_2$ of $T_{\text{scene}} \cup T_{\text{kernel}} \cup T_f \cup T_\Delta \cup I$ such that within $\mathcal{M}_1$ we have

$$\langle l_1, e_1 \rangle, \langle l_2, e_1 \rangle \in \Delta$$

$$\langle e_1, s_1 \rangle \in \text{part}$$

$$\langle r_1, s_1 \rangle, \langle r_2, s_1 \rangle \in \Delta$$

$$\langle r_3, s_1 \rangle, \langle r_4, s_1 \rangle \notin \Delta$$

and within $\mathcal{M}_2$ we have

$$\langle l_1, e_1 \rangle, \langle l_2, e_1 \rangle \in \Delta$$

$$\langle e_1, s_1 \rangle \in \text{part}$$

$$\langle r_1, s_1 \rangle, \langle r_2, s_1 \rangle \notin \Delta$$

$$\langle r_3, s_1 \rangle, \langle r_4, s_1 \rangle \in \Delta$$

□
It is important to note the difference between this depiction axiom and axioms 3.1 through 5.7 of $T_\Delta$. Those axioms depended on the existence of two distinct depicted edges: if we are given those edges, then the unique figure ground assignment of the surface entails the unique assignment of surface depiction literals to the regions of the image. With axiom 5.8, there is only one depicted edge, and we cannot directly appeal to the figure ground assignment. However, we can also think of this assignment as intuitively saying that the surface is on only one side of an edge; this notion of "side" is represented by the predicates inside and outside. Depiction axiom 5.8 is then simply a formalization of this intuition.

There are several comments to make about the depiction axioms $T_\Delta$ in Figure 5.18. Firstly, all depiction axioms for inside require that we know which lines depict edges in the same surface. As we will see in later chapters, the assertion that lines depict edges in the same surface is equivalent to grouping. In addition, there is a close relationship between the depiction axioms for inside and the assignment of surface depiction literals to regions in the image, which also plays a key role in grouping and algorithms that we will present based in grouping. Finally, because of the depiction of alignment relations, we will have two cases for the depiction inside for indeterminate edges.

### 5.5 Satisfiability of $T_\Delta$

**Theorem 5.4** Any structure in $\mathcal{M}_\Delta$ is a model of $\mathcal{T}_{scene} \cup \mathcal{T}_{kernel} \cup \mathcal{T}_{fg} \cup \mathcal{T}_\Delta$.

**Lemma 5.1** Let $\mathcal{M}$ be a structure in $\mathcal{M}_\Delta$. The depicted alignment graph which is a substructure of $\mathcal{M}$ satisfies axioms 5.1 - 5.4 of $T_\Delta$.

**Proof:** By property (1) in the definition of depicted alignment graphs, if there are any depicted edges, then there exists a nonempty depicted alignment graph.

Suppose for any variable assignment $\sigma$, $\mathcal{M}, \sigma \models \theta_i(a, l_1, l_2)$ iff the weight of the arc $(\sigma(l_1), \sigma(l_2))$ in the $\theta_i$ alignment graph in $\mathcal{M}$ is $\sigma(a)$, and suppose that $\mathcal{M}, \sigma \models \theta(a, e_1, e_2)$ iff the weight of the arc $(\sigma(e_1), \sigma(e_2))$ in the $\theta$ graph in $\mathcal{M}$ is $\sigma(a)$, such that $\sigma(a) \neq 0$.

By property (3) of the definition of depicted alignment graphs, if

$$\mathcal{M}, \sigma \models \Delta(l_1, e_1) \land \Delta(l_2, e_2) \land l_1 \neq l_2$$

then we also have

$$\mathcal{M}, \sigma \models \theta_i(\pi, l_1, l_2)$$

so that axiom 5.4 is satisfied.

By property (4) of the definition of depicted alignment graphs, if

$$\mathcal{M}, \sigma \models \Delta(l_1, e_1) \land \Delta(l_2, e_2) \land \neg \text{indet}(e_1, e_2) \land \theta(a, e_1, e_2)$$

then we also have

$$\mathcal{M}, \sigma \models \theta_i(a, l_1, l_2)$$

so that axiom 5.1 is satisfied.

By property (5) of the definition of depicted alignment graphs, if

$$\mathcal{M}, \sigma \models \Delta(l_1, e_1) \land \Delta(l_2, e_2) \land \text{indet}(e_1, e_2) \land \theta(a, e_1, e_2)$$
then we also have either
\[ \mathcal{M}, \sigma \models \theta_i(a, l_1, l_2) \]
or
\[ \mathcal{M}, \sigma \models \theta_i(a + \pi, l_1, l_2) \]
so that axiom 5.2 is satisfied.

By property (2) of the definition of depicted alignment subgraphs, if
\[ \mathcal{M}, \sigma \models \Delta(l_1, e_1) \land \Delta(l_2, e_2) \land \parallel(e_1, e_2) \]
then the weight of the arc \((l_1, l_2)\) in the depicted alignment subgraph is 0. By the definition of \(\theta_i\) graphs, we must have \((l_1, l_2) \in \parallel\) and hence
\[ \mathcal{M}, \sigma \models \parallel(l_1, l_2) \]
so that axiom 5.3 is satisfied. □

**Lemma 5.2** Let \(\mathcal{M}\) be a structure in \(\mathcal{M}^\Delta\). The depicted figure/ground graphs which are substructures of \(\mathcal{M}\) satisfy axiom 5.5 of \(T^\Delta\).

**Proof:** Let \(\mathcal{M}\) be a structure in \(\mathcal{M}^\Delta\).

\[ \mathcal{M} \models \]
\[ (\forall e_1, e_2, s, l_1, l_2, r) \) part\((e_1, s) \land \) part\((e_2, s) \land \Delta(l_1, e_1) \land \Delta(l_2, e_2) \land \text{in}(l_1, r) \land \lnot \text{indet}(e_1, e_2) \land e_1 \neq e_2 \supset \text{inside}(e_1, e_2) \equiv \text{between}(l_1, l_2, r) \]
is equivalent to the following: for any variable assignment \(\sigma\), if

\[ \langle \sigma(e_1), \sigma(s) \rangle, \langle \sigma(e_2), \sigma(s) \rangle \in \text{part} \]
\[ \langle \sigma(l_1), \sigma(e_1) \rangle, \langle \sigma(l_2), \sigma(e_2) \rangle \in \Delta \]
\[ \langle \sigma(e_1), \sigma(e_2) \rangle \notin \text{indet} \]
then
\[ \langle \sigma(e_1), \sigma(e_2) \rangle \in \text{inside} \]
iff
\[ \langle \sigma(r), \sigma(s) \rangle \in \Delta \text{ implies } \langle \sigma(l_1), \sigma(l_2), \sigma(r) \rangle \in \text{between}, \text{ and} \]
\[ \langle \sigma(l_1), \sigma(l_2), \sigma(r) \rangle \in \text{between}, \text{ implies } \langle \sigma(r), \sigma(s) \rangle \in \Delta. \]

By the definition of \(\mathcal{M}^\Delta\), there exists a unique depicted figure/ground graph for each depicted surface in \(\mathcal{M}\), and this graph is the sum of a unique depicted inside graph \(G\) containing \(\sigma(e_1), \sigma(e_2)\) (which are part of \(\sigma(s)\)) and a unique depicting between graph \(G'\) containing \(\sigma(l_1), \sigma(l_2)\) (which are in \(\sigma(r)\)).

There are four cases:

**Case 1:** \(\langle \sigma(e_1), \sigma(e_2) \rangle \in \text{inside}, \langle \sigma(r), \sigma(s) \rangle \in \Delta\)

Since \(\langle \sigma(e_1), \sigma(e_2) \rangle \in \text{inside}\), the definition of depicted inside graphs implies \(\sigma(e_1), \sigma(e_2) \in G\).

By Property (2) of the definition of a depicted figure/ground graph, we have \(\langle \sigma(l_1), \sigma(l_2) \rangle \in G'\).

By the definition of depicting between graphs, we get \(\langle \sigma(l_1), \sigma(l_2), \sigma(r) \rangle \in \text{between} \)

**Case 2:** \(\langle \sigma(e_1), \sigma(e_2) \rangle \in \text{inside}, \langle \sigma(r), \sigma(s) \rangle \notin \Delta\)

Since \(\langle \sigma(e_1), \sigma(e_2) \rangle \in \text{inside}\), the definition of depicted inside graphs implies \(\sigma(e_1), \sigma(e_2) \in G\).

By Property (3) of the definition of a depicted figure/ground graph, we have \(\langle \sigma(l_1), \sigma(l_2) \rangle \in G'\).

By the definition of depicting between graphs, we get \(\langle \sigma(l_1), \sigma(l_2), \sigma(r) \rangle \notin \text{between}\)

**Case 3:** \(\langle \sigma(r), \sigma(s) \rangle \in \Delta\)

Since \(\langle \sigma(r), \sigma(s) \rangle \in \Delta\), the definition of depicted inside graphs implies \(\sigma(r), \sigma(s) \in G\).

By Property (4) of the definition of a depicted figure/ground graph, we have \(\langle \sigma(l_1), \sigma(l_2) \rangle \in G'\).

By the definition of depicting between graphs, we get \(\langle \sigma(l_1), \sigma(l_2), \sigma(r) \rangle \in \text{between} \)

**Case 4:** \(\langle \sigma(r), \sigma(s) \rangle \notin \Delta\)

Since \(\langle \sigma(r), \sigma(s) \rangle \notin \Delta\), the definition of depicted inside graphs implies \(\sigma(r), \sigma(s) \notin G\).

By Property (5) of the definition of a depicted figure/ground graph, we have \(\langle \sigma(l_1), \sigma(l_2) \rangle \notin G'\).

By the definition of depicting between graphs, we get \(\langle \sigma(l_1), \sigma(l_2), \sigma(r) \rangle \notin \text{between} \)
Case 3: \( \langle \sigma(e_1), \sigma(e_2) \rangle \notin \text{inside}, \langle \sigma(r), \sigma(s) \rangle \in \Delta \)

Since \( \langle \sigma(e_1), \sigma(e_2) \rangle \notin \text{inside} \), the definition of depicted inside graphs implies \( \langle \sigma(e_1), \sigma(e_2) \rangle \notin G \).

By Property (2) of the definition of a depicted figure/ground graph, we have \( \langle \sigma(l_1), \sigma(l_2) \rangle \notin G' \).

By the definition of depicting between graphs, we get \( \langle \sigma(l_1), \sigma(l_2), \sigma(r) \rangle \notin \text{between} \)

Case 4: \( \langle \sigma(e_1), \sigma(e_2) \rangle \notin \text{inside}, \langle \sigma(r), \sigma(s) \rangle \notin \Delta \)

Since \( \langle \sigma(e_1), \sigma(e_2) \rangle \notin \text{inside} \), the definition of depicted inside graphs implies \( \langle \sigma(e_1), \sigma(e_2) \rangle \notin G \).

By Property (3) of the definition of a depicted figure/ground graph, we have \( \langle \sigma(l_1), \sigma(l_2) \rangle \in G' \).

By the definition of depicting between graphs, we get \( \langle \sigma(l_1), \sigma(l_2), \sigma(r) \rangle \in \text{between} \)

Thus axiom 5.5 of \( T_\Delta \) is satisfied.

\[ \square \]

Lemma 5.3 Let \( \mathcal{M} \) be a structure in \( \mathcal{M}_\Delta \). The depicted figure/ground graphs which are substructures of \( \mathcal{M} \) satisfy axiom 5.6 of \( T_\Delta \).

Proof: Let \( \mathcal{M} \) be a structure in \( \mathcal{M}_\Delta \).

\[ \mathcal{M} = (\forall e_1, e_2, s, l_1, l_2, r) \text{part}(e_1, s) \land \text{part}(e_2, s) \land \text{indet}(e_1, e_2) \land \Delta(l_1, e_1) \land \Delta(l_2, e_2) \land \theta_1(a, l_1, l_2) \land a < \pi \land \text{in}(l_1, r) \land e \neq e' \supset\]

\[ \text{convex}(e_1, e_2) \equiv \langle \Delta(r, s) \equiv \lnot \text{between}(l_1, l_2, r) \rangle \]

is equivalent to the following: for any variable assignment \( \sigma \), if

\[ \langle \sigma(e_1), \sigma(e_2) \rangle \in \text{indet} \]

\[ \langle \sigma(e_1), \sigma(s), \sigma(e_2), \sigma(s) \rangle \in \text{part} \]

\[ \langle \sigma(l_1), \sigma(e_1), \sigma(l_2), \sigma(e_2) \rangle \in \Delta \]

\[ \langle a, \sigma(l_1), \sigma(l_2) \rangle \in \theta_1, a < \pi \]

then

\[ \langle \sigma(e_1), \sigma(e_2) \rangle \in \text{convex} \]

iff

\[ \langle \sigma(r), \sigma(s) \rangle \in \Delta \text{ implies } \langle \sigma(l_1), \sigma(l_2), \sigma(r) \rangle \notin \text{between}, \]

\[ \langle \sigma(l_1), \sigma(l_2), \sigma(r) \rangle \notin \text{between}, \text{ implies } \langle \sigma(r), \sigma(s) \rangle \in \Delta. \]

By the definition of \( \mathcal{M}_\Delta \), there exists a unique depicted figure/ground graph for each depicted surface in \( \mathcal{M} \), and this graph is the sum of a unique depicted inside graph \( G \) containing \( \sigma(e_1), \sigma(e_2) \) (which are part of \( \sigma(s) \)) and a unique depicting between graph \( G' \) containing \( \sigma(l_1), \sigma(l_2) \) (which are in \( \sigma(r) \)).

There are four cases:

Case 1: \( \langle \sigma(e_1), \sigma(e_2) \rangle \in \text{convex}, \langle \sigma(r), \sigma(s) \rangle \in \Delta \)

By the definition of \( \text{convex} \), we have \( \langle \sigma(e_1), \sigma(e_2) \rangle \in \text{inside} \).

Since \( \langle \sigma(e_1), \sigma(e_2) \rangle \in \text{inside} \), the definition of depicted inside graphs implies \( \langle \sigma(e_1), \sigma(e_2) \rangle \in G \).

By Property (2) of the definition of a depicted figure/ground graph, we have \( \langle \sigma(l_1), \sigma(l_2) \rangle \in G' \).

By the definition of depicting between graphs, we get \( \langle \sigma(l_1), \sigma(l_2), \sigma(r) \rangle \in \text{between} \)

Case 2: \( \langle \sigma(e_1), \sigma(e_2) \rangle \in \text{convex}, \langle \sigma(r), \sigma(s) \rangle \notin \Delta \)

By the definition of \( \text{convex} \), we have \( \langle \sigma(e_1), \sigma(e_2) \rangle \in \text{inside} \).

Since \( \langle \sigma(e_1), \sigma(e_2) \rangle \in \text{inside} \), the definition of depicted inside graphs implies \( \langle \sigma(e_1), \sigma(e_2) \rangle \in G \).

By Property (2) of the definition of a depicted figure/ground graph, we have \( \langle \sigma(l_1), \sigma(l_2) \rangle \in G' \).

By the definition of depicting between graphs, we get \( \langle \sigma(l_1), \sigma(l_2), \sigma(r) \rangle \in \text{between} \)

Case 3: \( \langle \sigma(e_1), \sigma(e_2) \rangle \notin \text{inside}, \langle \sigma(r), \sigma(s) \rangle \in \Delta \)

By the definition of \( \text{convex} \), we have \( \langle \sigma(e_1), \sigma(e_2) \rangle \in \text{inside} \).

Since \( \langle \sigma(e_1), \sigma(e_2) \rangle \in \text{inside} \), the definition of depicted inside graphs implies \( \langle \sigma(e_1), \sigma(e_2) \rangle \in G \).

By Property (2) of the definition of a depicted figure/ground graph, we have \( \langle \sigma(l_1), \sigma(l_2) \rangle \in G' \).

By the definition of depicting between graphs, we get \( \langle \sigma(l_1), \sigma(l_2), \sigma(r) \rangle \in \text{between} \)

Case 4: \( \langle \sigma(e_1), \sigma(e_2) \rangle \notin \text{inside}, \langle \sigma(r), \sigma(s) \rangle \notin \Delta \)

By the definition of \( \text{convex} \), we have \( \langle \sigma(e_1), \sigma(e_2) \rangle \in \text{inside} \).

Since \( \langle \sigma(e_1), \sigma(e_2) \rangle \in \text{inside} \), the definition of depicted inside graphs implies \( \langle \sigma(e_1), \sigma(e_2) \rangle \in G \).
By Property (3) of the definition of a depicted figure/ground graph, we have \((\sigma(l_1), \sigma(l_2)) \in G\).
By the definition of depicting between graphs, we get \(\langle \sigma(l_1), \sigma(l_2), \sigma(r) \rangle \notin \text{between}\)

**Case 3:** \(\langle \sigma(e_1), \sigma(e_2) \rangle \notin \text{convex}, \langle \sigma(r), \sigma(s) \rangle \notin \Delta\)
By the definition of convex, we have \(\langle \sigma(e_1), \sigma(e_2) \rangle \in \text{inside}\).
Since \(\langle \sigma(e_1), \sigma(e_2) \rangle \notin \text{inside}\), the definition of depicted inside graphs implies \(\langle \sigma(e_1), \sigma(e_2) \rangle \notin G\).

By Property (2) of the definition of a depicted figure/ground graph, we have \(\langle \sigma(l_1), \sigma(l_2) \rangle \notin G\).

By the definition of depicting between graphs, we get \(\langle \sigma(l_1), \sigma(l_2), \sigma(r) \rangle \notin \text{between}\)

**Case 4:** \(\langle \sigma(e_1), \sigma(e_2) \rangle \notin \text{convex}, \langle \sigma(r), \sigma(s) \rangle \notin \Delta\)
By the definition of convex, we have \(\langle \sigma(e_1), \sigma(e_2) \rangle \in \text{inside}\).
Since \(\langle \sigma(e_1), \sigma(e_2) \rangle \notin \text{inside}\), the definition of depicted inside graphs implies \(\langle \sigma(e_1), \sigma(e_2) \rangle \notin G\).

By Property (3) of the definition of a depicted figure/ground graph, we have \(\langle \sigma(l_1), \sigma(l_2) \rangle \notin G\).

By the definition of depicting between graphs, we get \(\langle \sigma(l_1), \sigma(l_2), \sigma(r) \rangle \notin \text{between}\)

Thus axiom 5.6 of \(T_\Delta\) is satisfied. □

**Lemma 5.4** Let \(\mathcal{M}\) be a structure in \(\mathcal{M}^\Delta\). The depicted figure/ground graphs which are substructures of \(\mathcal{M}\) satisfy axiom 5.6 of \(T_\Delta\).

**Proof:** Let \(\mathcal{M}\) be a structure in \(\mathcal{M}^\Delta\).
\[
\mathcal{M} \models \langle v e_1, e_2, s, l_1, l_2, r \rangle \text{part}(e_1, s) \land \text{part}(e_2, s) \land \text{indet}(e_1, e_2) \land \Delta(l_1, e_1) \land \Delta(l_2, e_2) \\
\land \theta_1(a, l_1, l_2) \land a > \pi \land \text{in}(l_1, r) \land e_1 \neq e_2 \supset \\
\langle \text{convex}(e_1, e_2) \equiv (\Delta(r, s) \equiv \text{between}(l_1, l_2, r)) \rangle
\]
is equivalent to the following: for any variable assignment \(\sigma\), if
\[
\langle \sigma(e_1), \sigma(e_2) \rangle \in \text{indet} \\
\langle \sigma(e_1), \sigma(s) \rangle, \langle \sigma(e_2), \sigma(s) \rangle \in \text{part} \\
\langle \sigma(l_1), \sigma(e_1) \rangle, \langle \sigma(l_2), \sigma(e_2) \rangle \in \Delta \\
\langle a, \sigma(l_1), \sigma(l_2) \rangle \in \theta_1, a < \pi
\]
then
\[
\langle \sigma(e_1), \sigma(e_2) \rangle \in \text{convex}
\]
iff
\[
\langle \sigma(r), \sigma(s) \rangle \in \Delta \text{ implies } \langle \sigma(l_1), \sigma(l_2), \sigma(r) \rangle \in \text{between}, \text{ and} \\
\langle \sigma(l_1), \sigma(l_2), \sigma(r) \rangle \notin \text{between}, \text{ implies } \langle \sigma(r), \sigma(s) \rangle \in \Delta.
\]
By the definition of \(\mathcal{M}^\Delta\), there exists a unique depicted figure/ground graph for each depicted surface in \(\mathcal{M}\), and this graph is the sum of a unique depicted inside graph \(G\) containing \(\sigma(e_1), \sigma(e_2)\) (which are part of \(\sigma(s)\)) and a unique depicting between graph \(G'\) containing \(\sigma(l_1), \sigma(l_2)\) (which are in \(\sigma(r)\)).

There are four cases:

**Case 1:** \(\langle \sigma(e_1), \sigma(e_2) \rangle \in \text{convex}, \langle \sigma(r), \sigma(s) \rangle \in \Delta\)
By the definition of convex, we have \(\langle \sigma(e_1), \sigma(e_2) \rangle \in \text{inside}\).
Since \(\langle \sigma(e_1), \sigma(e_2) \rangle \in \text{inside}\), the definition of depicted inside graphs implies \(\langle \sigma(e_1), \sigma(e_2) \rangle \in G\).

By Property (2) of the definition of a depicted figure/ground graph, we have \(\langle \sigma(l_1), \sigma(l_2) \rangle \in G'\).

By the definition of depicting between graphs, we get \(\langle \sigma(l_1), \sigma(l_2), \sigma(r) \rangle \in \text{between}\)
Case 2: \( \langle \sigma(e_1), \sigma(e_2) \rangle \in \text{convex}, \langle \sigma(r), \sigma(s) \rangle \notin \Delta \)

By the definition of \text{convex}, we have \( \langle \sigma(e_1), \sigma(e_2) \rangle \in \text{inside} \).

Since \( \langle \sigma(e_1), \sigma(e_2) \rangle \in \text{inside} \), the definition of depicted inside graphs implies \( \langle \sigma(e_1), \sigma(e_2) \rangle \in G \).

By Property (3) of the definition of a depicted figure/ground graph, we have \( \langle \sigma(l_1), \sigma(l_2) \rangle \in G' \).

By the definition of depicting between graphs, we get \( \langle \sigma(l_1), \sigma(l_2), \sigma(r) \rangle \notin \text{between} \).

Case 3: \( \langle \sigma(e_1), \sigma(e_2) \rangle \notin \text{convex}, \langle \sigma(r), \sigma(s) \rangle \notin \Delta \)

By the definition of \text{convex}, we have \( \langle \sigma(e_1), \sigma(e_2) \rangle \in \text{inside} \).

Since \( \langle \sigma(e_1), \sigma(e_2) \rangle \notin \text{inside} \), the definition of depicted inside graphs implies \( \langle \sigma(e_1), \sigma(e_2) \rangle \in G \).

By Property (2) of the definition of a depicted figure/ground graph, we have \( \langle \sigma(l_1), \sigma(l_2) \rangle \notin G' \).

By the definition of depicting between graphs, we get \( \langle \sigma(l_1), \sigma(l_2), \sigma(r) \rangle \notin \text{between} \).

Case 4: \( \langle \sigma(e_1), \sigma(e_2) \rangle \notin \text{convex}, \langle \sigma(r), \sigma(s) \rangle \notin \Delta \)

By the definition of \text{convex}, we have \( \langle \sigma(e_1), \sigma(e_2) \rangle \in \text{inside} \).

Since \( \langle \sigma(e_1), \sigma(e_2) \rangle \notin \text{inside} \), the definition of depicted inside graphs implies \( \langle \sigma(e_1), \sigma(e_2) \rangle \notin G \).

By Property (3) of the definition of a depicted figure/ground graph, we have \( \langle \sigma(l_1), \sigma(l_2) \rangle \in G' \).

By the definition of depicting between graphs, we get \( \langle \sigma(l_1), \sigma(l_2), \sigma(r) \rangle \in \text{between} \).

Thus axiom 5.7 of \( T_\Delta \) is satisfied. \( \square \)

**Lemma 5.5** Let \( M \) be a structure in \( M^\Delta \). The solitary depicting graphs which are substructures of \( M \) satisfy axiom 5.8 of \( T_\Delta \).

**Proof:** Let \( M \) be a structure in \( M^\Delta \).

\[
M \models (\forall l_1, l_2, r_1, r_2, e, s) \in(l_1, r_1) \land \in(l_2, r_2) \land \Delta(l_1, e) \land \Delta(l_2, e) \land \text{part}(e, s) \supset \\
[\langle \Delta(r_1, s) \equiv \Delta(r_2, s) \rangle \equiv \langle \text{between}(l_1, l_2, r_1) \equiv \neg \text{between}(l_2, l_1, r_2) \rangle]
\]

is equivalent to saying that for any variable assignment \( \sigma \),

\[
M, \sigma \models \Delta(r_1, s) \equiv \Delta(r_2, s)
\]

iff

\[
M, \sigma \models \text{between}(l_1, l_2, r_1) \equiv \neg \text{between}(l_2, l_1, r_2)
\]

This follows from the definition of solitary depicting graphs, since all regions depict the same surface iff the subgraph of the depicting between graph is a directed acyclic graph, and

\[
M, \sigma \models \text{between}(l_1, l_2, r_1) \equiv \neg \text{between}(l_2, l_1, r_2)
\]

iff the depicting between graph is acyclic. \( \square \)

### 5.6 Characterizing the Models of \( T_\Delta \)

The depiction axioms must preserve the properties of the figure ground axioms presented in the previous section. We thus present three theorems which demonstrate this relationship between surfaces and depicted surfaces. The first shows that the assignment of figure ground predicates for a depicted surface is preserved by depiction. The second shows that since depicted surfaces can be given a unique assignment of figure ground predicates, the depiction axioms can be used to uniquely assign surface depiction literals to regions. The third theorem shows that these properties also hold when only one edge in a surface is
depicted. Again, all of these theorems can be considered to be relativized completeness theorems, since they state the conditions under which $T_{\text{scene}} \cup T_{\text{kernel}} \cup T_{fg} \cup T_{\Delta}$ completely specifies the extensions of inside and outside for depicted surfaces.

**Theorem 5.5** Let $\mathcal{M}, \mathcal{M}'$ be models of $T_{\text{scene}} \cup T_{\text{kernel}} \cup T_{fg} \cup T_{\Delta}$ with the same domain and which agree on the extension of indet. Suppose for any variable assignment $\sigma$.

\[
\mathcal{M}, \sigma \models \Delta(l, e) \land \Delta(l', e') \land \Delta(r, s) \land \text{part}(e, s) \land \text{part}(e', s) \land e \neq e' \land \text{in}(l, r)
\]

iff

\[
\mathcal{M}', \sigma \models \Delta(l, e) \land \Delta(l', e') \land \Delta(r, s) \land \text{part}(e, s) \land \text{part}(e', s) \land e \neq e' \land \text{in}(l, r)
\]

Then $\mathcal{M}, \mathcal{M}'$ agree on the extensions of inside and outside.

**Proof:** Let $\mathcal{M}, \mathcal{M}'$ be two models of $T_{\text{scene}} \cup T_{\text{kernel}} \cup T_{fg} \cup T_{\Delta}$ such that for any variable assignment $\sigma$ we have

\[
\mathcal{M}, \sigma \models \Delta(l, e) \land \Delta(l', e') \land \Delta(r, s) \land \text{part}(e, s) \land \text{part}(e', s) \land e \neq e' \land \text{in}(l, r)
\]

iff

\[
\mathcal{M}', \sigma \models \Delta(l, e) \land \Delta(l', e') \land \Delta(r, s) \land \text{part}(e, s) \land \text{part}(e', s) \land e \neq e' \land \text{in}(l, r)
\]

There are two cases depending on the extension of the indet relation:

**Case 1:** $\mathcal{M}, \sigma \models \neg \text{indet}(e, e')$ and $\mathcal{M}', \sigma \models \neg \text{indet}(e, e')$ There are two subcases depending on whether or not the region $r$ is between the lines $l$ and $l'$:

1. If $\mathcal{M}, \sigma \models \text{between}(l, l', r)$, then by axiom 5.5 of $T_{\Delta}$ and axiom 4.13 of $T_{fg}$, we have

\[
\mathcal{M}, \sigma \models \text{inside}(e, e') \land \neg \text{outside}(e, e')
\]

and

\[
\mathcal{M}', \sigma \models \text{inside}(e, e') \land \neg \text{outside}(e, e')
\]

2. If $\mathcal{M}, \sigma \models \neg \text{between}(l, l', r)$, then by axiom 5.5 of $T_{\Delta}$ and axiom 4.13 of $T_{fg}$, we have

\[
\mathcal{M}, \sigma \models \neg \text{inside}(e, e') \land \text{outside}(e, e')
\]

and

\[
\mathcal{M}', \sigma \models \neg \text{inside}(e, e') \land \text{outside}(e, e')
\]

**Case 2:** $\mathcal{M}, \sigma \models \text{indet}(e, e')$ and $\mathcal{M}', \sigma \models \text{indet}(e, e')$ By axiom (2) of $T_{\Delta}$, there are two subcases depending on the relative alignment of the lines $l$ and $l'$:

1. If $\mathcal{M}, \sigma \models \theta_l(a, l, l') \land a < \pi$, then there are two subcases depending on whether or not the region $r$ is between the lines $l$ and $l'$:

   (a) If $\mathcal{M}, \sigma \models \text{between}(l, l', r)$, then by axiom 5.6 of $T_{\Delta}$ and axiom 4.13 of $T_{fg}$, we have

   \[
   \mathcal{M}, \sigma \models \text{convex}(e, e')
   \]

   iff

   \[
   \mathcal{M}', \sigma \models \text{convex}(e, e')
   \]

   from which it follows

   \[
   \mathcal{M}, \sigma \models \text{inside}(e, e') \land \text{outside}(e, e')
   \]

   and

   \[
   \mathcal{M}', \sigma \models \text{inside}(e, e') \land \text{outside}(e, e')
   \]


(b) If $\mathcal{M}, \sigma \models \neg \text{between}(l, l', r)$, then by axiom 5.6 of $T_{\Delta}$ and axiom 4.13 of $T_{tg}$, we have

$$\mathcal{M}, \sigma \models \neg \text{convex}(e, e')$$

iff

$$\mathcal{M}', \sigma \models \neg \text{convex}(e, e')$$

from which it follows

$$\mathcal{M}, \sigma \models \neg \text{inside}(e, e') \land \neg \text{outside}(e, e')$$

and

$$\mathcal{M}', \sigma \models \neg \text{inside}(e, e') \land \neg \text{outside}(e, e')$$

2. If $\mathcal{M}, \sigma \models \theta_1(a, l, l') \land a > \pi$, then there are two subcases, depending on whether or not the region $r$ is between the lines $l$ and $l'$:

(a) If $\mathcal{M}, \sigma \models \text{between}(l, l', r)$, then by axiom 5.7 of $T_{\Delta}$ and axiom 4.13 of $T_{tg}$, we have

$$\mathcal{M}, \sigma \models \neg \text{convex}(e, e')$$

iff

$$\mathcal{M}', \sigma \models \neg \text{convex}(e, e')$$

from which it follows

$$\mathcal{M}, \sigma \models \text{inside}(e, e') \land \text{outside}(e, e')$$

and

$$\mathcal{M}', \sigma \models \text{inside}(e, e') \land \text{outside}(e, e')$$

(b) If $\mathcal{M}, \sigma \models \neg \text{between}(l, l', r)$, then by axiom 5.7 of $T_{\Delta}$ and axiom 4.13 of $T_{tg}$, we have

$$\mathcal{M}, \sigma \models \text{convex}(e, e')$$

iff

$$\mathcal{M}', \sigma \models \text{convex}(e, e')$$

from which it follows

$$\mathcal{M}, \sigma \models \neg \text{inside}(e, e') \land \neg \text{outside}(e, e')$$

and

$$\mathcal{M}', \sigma \models \neg \text{inside}(e, e') \land \neg \text{outside}(e, e')$$

Thus in all cases, $\mathcal{M}$ and $\mathcal{M}'$ agree on the extensions of inside and outside. □

**Theorem 5.6** Let $\mathcal{M}, \mathcal{M}'$ be models of $T_{\text{scene}} \cup T_{\text{kernel}} \cup T_{tg} \cup T_{\Delta}$ with the same domain and which agree on the extension of indet. Suppose for any variable assignment $\sigma$,

$$\mathcal{M}, \sigma \models \Delta(l, e) \land \Delta(l', e') \land \neg \Delta(r, s) \land \text{part}(e, s) \land \text{part}(e', s) \land e \neq e' \land \text{in}(l, r)$$

iff

$$\mathcal{M}', \sigma \models \Delta(l, e) \land \Delta(l', e') \land \neg \Delta(r, s) \land \text{part}(e, s) \land \text{part}(e', s) \land e \neq e' \land \text{in}(l, r)$$

Then $\mathcal{M}, \mathcal{M}'$ agree on the extensions of inside and outside.

**Proof:** Let $\mathcal{M}, \mathcal{M}'$ be two models of $T_{\text{scene}} \cup T_{\text{kernel}} \cup T_{tg} \cup T_{\Delta}$ such that for any variable assignment $\sigma$ we have

$$\mathcal{M}, \sigma \models \Delta(l, e) \land \Delta(l', e') \land \neg \Delta(r, s) \land \text{part}(e, s) \land \text{part}(e', s) \land e \neq e' \land \text{in}(l, r)$$

iff

$$\mathcal{M}', \sigma \models \Delta(l, e) \land \Delta(l', e') \land \neg \Delta(r, s) \land \text{part}(e, s) \land \text{part}(e', s) \land e \neq e' \land \text{in}(l, r)$$

There are two cases depending on the extension of the indet relation:

**Case 1:** $\mathcal{M}, \sigma \models \neg \text{indet}(e, e')$ and $\mathcal{M}', \sigma \models \neg \text{indet}(e, e')$ There are two subcases depending on whether or not the region $r$ is between the lines $l$ and $l'$:
1. If $\mathcal{M}, \sigma \models \text{between}(l, l', r)$, then by axiom 5.5 of $T_{\Delta}$ and axiom 4.13 of $T_{fg}$, we have
$$\mathcal{M}, \sigma \models \neg \text{inside}(e, e') \land \text{outside}(e, e')$$
and
$$\mathcal{M}', \sigma \models \neg \text{inside}(e, e') \land \text{outside}(e, e')$$

2. If $\mathcal{M}, \sigma \models \neg \text{between}(l, l', r)$, then by axiom 5.5 of $T_{\Delta}$ and axiom 4.13 of $T_{fg}$, we have
$$\mathcal{M}, \sigma \models \text{inside}(e, e') \land \neg \text{outside}(e, e')$$
and
$$\mathcal{M}', \sigma \models \text{inside}(e, e') \land \neg \text{outside}(e, e')$$

Case 2: $\mathcal{M}, \sigma \models \text{indet}(e, e')$ and $\mathcal{M}', \sigma \models \text{indet}(e, e')$ By axiom (2) of $T_{\Delta}$, there are two subcases depending on the relative alignment of the lines $l$ and $l'$:

1. If $\mathcal{M}, \sigma \models \theta_l(1, a, l, l') \land a < \pi$, then there are two subcases depending on whether or not the region $r$ is between the lines $l$ and $l'$:

(a) If $\mathcal{M}, \sigma \models \text{between}(l, l', r)$, then by axiom 5.6 of $T_{\Delta}$ and axiom 4.13 of $T_{fg}$, we have
$$\mathcal{M}, \sigma \models \neg \text{convex}(e, e')$$
iff
$$\mathcal{M}', \sigma \models \neg \text{convex}(e, e')$$
from which it follows
$$\mathcal{M}, \sigma \models \neg \text{inside}(e, e') \land \neg \text{outside}(e, e')$$
and
$$\mathcal{M}', \sigma \models \neg \text{inside}(e, e') \land \neg \text{outside}(e, e')$$

(b) If $\mathcal{M}, \sigma \models \neg \text{between}(l, l', r)$, then by axiom 5.6 of $T_{\Delta}$ and axiom 4.13 of $T_{fg}$, we have
$$\mathcal{M}, \sigma \models \text{convex}(e, e')$$
iff
$$\mathcal{M}', \sigma \models \text{convex}(e, e')$$
from which it follows
$$\mathcal{M}, \sigma \models \text{inside}(e, e') \land \text{outside}(e, e')$$
and
$$\mathcal{M}', \sigma \models \text{inside}(e, e') \land \text{outside}(e, e')$$

2. If $\mathcal{M}, \sigma \models \theta_l(1, a, l, l') \land a > \pi$, then there are two subcases, depending on whether or not the region $r$ is between the lines $l$ and $l'$:

(a) If $\mathcal{M}, \sigma \models \text{between}(l, l', r)$, then by axiom 5.7 of $T_{\Delta}$ and axiom 4.13 of $T_{fg}$, we have
$$\mathcal{M}, \sigma \models \text{convex}(e, e')$$
iff
$$\mathcal{M}', \sigma \models \text{convex}(e, e')$$
from which it follows
$$\mathcal{M}, \sigma \models \neg \text{inside}(e, e') \land \neg \text{outside}(e, e')$$
and
$$\mathcal{M}', \sigma \models \neg \text{inside}(e, e') \land \neg \text{outside}(e, e')$$
(b) If $\mathcal{M}, \sigma \models \neg\text{between}(l, l', r)$, then by axiom 5.7 of $T_\Delta$ and axiom 4.13 of $T_{fg}$, we have

$\mathcal{M}, \sigma \models \neg\text{convex}(e, e')$

iff

$\mathcal{M}', \sigma \models \neg\text{convex}(e, e')$

from which it follows

$\mathcal{M}, \sigma \models \text{inside}(e, e') \land \text{outside}(e, e')$

and

$\mathcal{M}', \sigma \models \text{inside}(e, e') \land \text{outside}(e, e')$

Thus in all cases, $\mathcal{M}$ and $\mathcal{M}'$ agree on the extensions of $\text{inside}$ and $\text{outside}$. □

**Corollary 5.1** Let $\mathcal{M}$ be a model of $T_\Delta \cup T_{fg} \cup T_{kernel} \cup T_{scene}$. Then $\mathcal{M}$ contains a depicted figure/ground graph for each depicted surface in $\mathcal{M}$ that contains multiple depicted edges.

**Proof:** We will construct a depicted inside graph for each depicted surface, and a depicting between graph for each region which depicts one of these surfaces. We will then show that the sum of these graphs is isomorphic to a depicted figure/ground graph.

We will use the proofs for Theorems 5.5 and 5.6, in which we characterized the extension of $\text{inside}$. Construct a graph $G = (N, A)$ in which $N$ is the set of depicted edges in $\mathcal{M}$, and $(e_1, e_2) \in A$ iff $(e_1, e_2) \in \text{inside}$.

In cases 1.1 and 1.2 in the proof of Theorem 5.5, we have $\langle r, s \rangle \in \Delta$, $(e_1, e_2) \notin \text{indet}$, and

$(e_1, e_2) \in A$ iff $\langle l_1, l_2 \rangle \in \text{between}$

In cases 2.1(a) and 2.1(b) in the proof of Theorem 5.5, we have $\langle r, s \rangle \in \Delta$, $(e_1, e_2) \in \text{indet}$, and

$(e_1, e_2) \in A$ iff $\langle l_1, l_2 \rangle \in \text{between}$

In cases 2.2(a) and 2.2(b) in the proof of Theorem 5.5, we have $\langle r, s \rangle \in \Delta$, $(e_1, e_2) \in \text{indet}$, and

$(e_1, e_2) \in A$ iff $\langle l_1, l_2 \rangle \in \text{between}$

By Theorem 5.5, these are the only possibilities for the extension of $\text{inside}$ given that the region $r$ depicts the surface $s$, so conditions (2), (4), and (6) in the definition of depicted figure/ground graphs are satisfied.

In cases 1.1 and 1.2 in the proof of Theorem 5.6, we have $\langle r, s \rangle \in \Delta$, $(e_1, e_2) \notin \text{indet}$, and

$(e_1, e_2) \in A$ iff $\langle l_1, l_2 \rangle \notin \text{between}$

In cases 2.1(a) and 2.1(b) in the proof of Theorem 5.6, we have $\langle r, s \rangle \in \Delta$, $(e_1, e_2) \in \text{indet}$, and

$(e_1, e_2) \in A$ iff $\langle l_1, l_2 \rangle \notin \text{between}$

In cases 2.2(a) and 2.2(b) in the proof of Theorem 5.6, we have $\langle r, s \rangle \in \Delta$, $(e_1, e_2) \in \text{indet}$, and

$(e_1, e_2) \in A$ iff $\langle l_1, l_2 \rangle \notin \text{between}$

By Theorem 5.6, these are the only possibilities for the extension of $\text{inside}$ given that the region $r$ does not depict the surface $s$, so conditions (3), (5) and (7) in the definition of depicted figure/ground graphs are satisfied.

Thus, the graph which we have constructed is indeed isomorphic to a depicted figure/ground graph. □
Theorem 5.7 Let \( M, M' \) be models of \( T_{\text{scene}} \cup T_{\text{kernel}} \cup T_{f_9} \cup T_{\Delta} \) with the same domain which agree on the extension of \( \text{indet} \). Suppose for any variable assignment \( \sigma \),
\[
M, \sigma \models \Delta(l, e) \land \Delta(l', e') \land \text{part}(e, s) \land \text{part}(e', s) \land e \neq e' \land \text{in}(l, r)
\]
iff
\[
M', \sigma \models \Delta(l, e) \land \Delta(l', e') \land \text{part}(e, s) \land \text{part}(e', s) \land e \neq e' \land \text{in}(l, r)
\]
and suppose \( M, M' \) agree on the extensions of \( \text{inside} \) and \( \text{outside} \). Then \( M, M' \) agree on the extension of \( \Delta \) for regions and surfaces.

Proof: There are two major cases – \( \text{indet}(e, e') \) and \( \neg\text{indet}(e, e') \). First suppose
\[
M, \sigma \models \neg\text{indet}(e, e'), M', \sigma \models \neg\text{indet}(e, e')
\]

By axiom 4.13 of \( T_{f_9} \), we have two cases:

1. If
\[
M, \sigma \models \text{inside}(e, e') \land \neg\text{outside}(e, e')
\]

there are two cases:

(a) If \( M, \sigma \models \text{between}(l, l', r) \), then axiom 5.5 of \( T_{\Delta} \) gives
\[
M, \sigma \models \neg\Delta(r, s)
\]
\[
M', \sigma \models \neg\Delta(r, s)
\]

(b) If \( M, \sigma \models \neg\text{between}(l, l', r) \), then axiom 5.5 of \( T_{\Delta} \) gives
\[
M, \sigma \models \Delta(r, s)
\]
\[
M', \sigma \models \Delta(r, s)
\]

2. If
\[
M, \sigma \models \neg\text{inside}(e, e') \land \text{outside}(e, e')
\]

there are two cases:

(a) If \( M, \sigma \models \neg\text{between}(l, l', r) \), then axiom 5.5 of \( T_{\Delta} \) gives
\[
M, \sigma \models \neg\Delta(r, s)
\]
\[
M', \sigma \models \neg\Delta(r, s)
\]

(b) If \( M, \sigma \models \text{between}(l, l', r) \), then axiom 5.5 of \( T_{\Delta} \) gives
\[
M, \sigma \models \Delta(r, s)
\]
\[
M', \sigma \models \Delta(r, s)
\]

Next, if
\[
M, \sigma \models \text{indet}(e, e'),
\]
\[
M', \sigma \models \text{indet}(e, e')
\]

then by axiom (2) of \( T_{\Delta} \) there are two cases:

1. If \( M, \sigma \models \theta_i(a, l, l') \land a < \pi \), then by axiom 4.13 of \( T_{f_9} \) there are two cases
(a) If
\[ M, \sigma \models inside(e, e') \land outside(e, e') \]
\[ M', \sigma \models inside(e, e') \land outside(e, e') \]
then there are two cases:
  i. If \( M, \sigma \models between(l, l', r) \) then axiom 5.6 of \( T_\Delta \) gives
  \[ M, \sigma \models \neg \Delta(r, s) \]
  \[ M', \sigma \models \neg \Delta(r, s) \]
  ii. If \( M, \sigma \models \neg between(l, l', r) \) then axiom 5.6 of \( T_\Delta \) gives
  \[ M, \sigma \models \Delta(r, s) \]
  \[ M', \sigma \models \Delta(r, s) \]

(b) If
\[ M, \sigma \models \neg inside(e, e') \land \neg outside(e, e') \]
\[ M', \sigma \models \neg inside(e, e') \land \neg outside(e, e') \]
then there are two cases:
  i. If \( M, \sigma \models between(l, l', r) \) then axiom 5.6 of \( T_\Delta \) gives
  \[ M, \sigma \models \Delta(r, s) \]
  \[ M', \sigma \models \Delta(r, s) \]
  ii. If \( M, \sigma \models \neg between(l, l', r) \) then axiom 5.6 of \( T_\Delta \) gives
  \[ M, \sigma \models \neg \Delta(r, s) \]
  \[ M', \sigma \models \neg \Delta(r, s) \]

(c) If \( M, \sigma \models \theta(a, l, l') \land a > \pi \), then by axiom 4.13 of \( T_\theta \) there are two cases
  i. If
  \[ M, \sigma \models inside(e, e') \land outside(e, e') \]
  \[ M', \sigma \models inside(e, e') \land outside(e, e') \]
  then there are two cases:
  A. If \( M, \sigma \models between(l, l', r) \) then axiom 5.7 of \( T_\Delta \) gives
  \[ M, \sigma \models \Delta(r, s) \]
  \[ M', \sigma \models \Delta(r, s) \]
  B. If \( M, \sigma \models \neg between(l, l', r) \) then axiom 5.7 of \( T_\Delta \) gives
  \[ M, \sigma \models \neg \Delta(r, s) \]
  \[ M', \sigma \models \neg \Delta(r, s) \]
  ii. If
  \[ M, \sigma \models \neg inside(e, e') \land \neg outside(e, e') \]
  \[ M', \sigma \models \neg inside(e, e') \land \neg outside(e, e') \]
  then there are two cases:
A. If \( M, \sigma \models \text{between}(l, l', r) \) then axiom 5.7 of \( T_\Delta \) gives
\[
M, \sigma \models \neg \Delta(r, s)
\]
\[
M', \sigma \models \neg \Delta(r, s)
\]

B. If \( M, \sigma \models \neg \text{between}(l, l', r) \) then axiom 5.7 of \( T_\Delta \) gives
\[
M, \sigma \models \Delta(r, s)
\]
\[
M', \sigma \models \Delta(r, s)
\]

In all cases, \( M \) and \( M' \) agree on the extension of \( \Delta \) for regions and surfaces. \( \square \)

This theorem states that if we know that two lines depict edges in the same surface, and we know the figure ground relations of these depicted edges, then we can uniquely assign surface depiction literals to the regions.

**Theorem 5.8** Let \( M, M' \) be models of \( T_{\text{scene}} \cup T_{\text{kernel}} \cup T_{fg} \cup T_\Delta \) with the same domain. Suppose for any variable assignment \( \sigma \),
\[
M, \sigma \models \text{in}(l, r) \land \text{in}(l, r'') \land \text{in}(l', r') \land \text{in}(l'', r''')
\]
\[
\land \text{between}(l, l', r) \land \neg \text{between}(l, l', r'') \land \neg \text{between}(l', l, r') \land \text{between}(l', l, r''')
\]
iff
\[
M', \sigma \models \text{in}(l, r) \land \text{in}(l, r'') \land \text{in}(l', r') \land \text{in}(l'', r''')
\]
\[
\land \text{between}(l, l', r) \land \neg \text{between}(l, l', r'') \land \neg \text{between}(l', l, r') \land \text{between}(l', l, r''')
\]
Suppose also that
\[
M, \sigma \models \Delta(l, e) \land \Delta(l', e') \land \text{part}(e, s)
\]
iff
\[
M', \sigma \models \Delta(l, e) \land \Delta(l', e') \land \text{part}(e, s)
\]
Then \( M, M' \) agree on the extension of \( \Delta \) for regions and surfaces.

**Proof:** By the premises of this theorem, the models \( M \) and \( M' \) agree on the extension of \( \text{between} \) for some tuple of lines and regions, and they also agree on the extension of \( \Delta \) for these lines.

The theorem follows easily from axiom 5.8 of \( T_\Delta \):

We have either
\[
M, \sigma \models \Delta(r, s) \land \neg \Delta(r'', s) \land \Delta(r', s) \land \neg \Delta(r''', s)
\]
and
\[
M', \sigma \models \Delta(r, s) \land \neg \Delta(r'', s) \land \Delta(r', s) \land \neg \Delta(r''', s)
\]
or we have
\[
M, \sigma \models \neg \Delta(r, s) \land \Delta(r'', s) \land \neg \Delta(r', s) \land \Delta(r''', s)
\]
and
\[
M', \sigma \models \neg \Delta(r, s) \land \Delta(r'', s) \land \neg \Delta(r', s) \land \Delta(r''', s)
\]
\( \square \)

**Lemma 5.6** Any model of \( T_{\text{scene}} \cup T_{\text{kernel}} \cup T_{fg} \cup T_\Delta \) contains a solitary depicting graph for each depicted surface that has a unique depicted edge in the model.
Proof: Let $\mathcal{M}$ be a model of $T_{\text{scene}} \cup T_{\text{kernel}} \cup T_{fg} \cup T_{\Delta}$. By axiom 5.8 of $T_{\Delta}$, we have

$$\mathcal{M}, \sigma \vDash \Delta(r_1, s) \equiv \Delta(r_2, s)$$

iff

$$\mathcal{M}, \sigma \vDash \text{between}(l_1, l_2, r_1) \equiv \neg \text{between}(l_2, l_1, r_2)$$

Construct a graph $G = (N, A)$ such that $N$ is the set of lines depicting edges in $\sigma(s)$, and $(x, y) \in A$ iff $(x, y, r) \in \text{between}$, where $r$ is a region depicting $\sigma(s)$ in $\mathcal{M}$. It is easy to see that $G$ is a subgraph of the depicting between graph.

Suppose $\sigma(r_1), \sigma(r_2)$ both depict the surface $\sigma(s)$. Then we have

$$\mathcal{M}, \sigma \vDash \text{between}(l_1, l_2, r_1) \lor \text{between}(l_2, l_1, r_2)$$

so that $G$ is connected, and we also have

$$\mathcal{M}, \sigma \vDash \text{between}(l_1, l_2, r_1) \supset \neg \text{between}(l_2, l_1, r_2)$$

so that $G$ is acyclic.

The transitivity of $G$ is a consequence of the following sentence, which holds in all structures in $\mathcal{M}^\text{tuple}$:

$$\mathcal{M} \models (\forall l_1, l_2, l_3, r_1, r_2, r_3) \text{between}(l_1, l_2, r_1) \land \text{between}(l_2, l_3, r_2) \land \theta_i(\pi, l_1, l_2) \land \theta_i(\pi, l_2, l_3) \supset \text{between}(l_1, l_3, r_2) \land \theta_i(\pi, l_1, l_3)$$

Since $G$ exists iff the regions depict the same surface, $G$ is a solitary depicting graph. □

Lemma 5.7 Any model of $T_{\text{scene}} \cup T_{\text{kernel}} \cup T_{fg} \cup T_{\Delta}$ contains a unique depicted alignment graph for each depicted surface in the model.

Proof: Consider a model $\mathcal{M}$ of $T_{\text{scene}} \cup T_{\text{kernel}} \cup T_{fg} \cup T_{\Delta}$ containing two lines $l_1, l_2$ which both depict edges some surface $s \in M$. There will thus be a depicted $\theta_i$ subgraph for $s$. There are two cases – either they depict distinct edges in the surface, or they depict the same edge in the surface.

Suppose that the lines depict distinct edges, so that

$$\langle l_1, e_1 \rangle, \langle l_2, e_2 \rangle \in \Delta$$

$$\langle e_1, s \rangle, \langle e_2, s \rangle \in \text{part}$$

By axiom 4.6 of $T_{fg}$, there are three subcases – either $\langle e_1, e_2 \rangle \in \text{indet}$ or $\langle e_1, e_2 \rangle \notin \text{indet}$, or $\langle e_1, e_2 \rangle \in \text{parallel}$.

In the first subcase, axiom 5.1 of $T_{\Delta}$ gives

$$\langle a, e_1, e_2 \rangle \in \theta \iff \langle a, l_1, l_2 \rangle \in \theta_i$$

Thus, the weight of the arc in the depicting $\theta_i$ graph is equal to the relative alignment of the edges, satisfying condition (4) in the definition of depicted alignment graphs.

In the second subcase, axiom 5.2 of $T_{\Delta}$ gives $\langle a, e_1, e_2 \rangle \in \theta$ iff either $\langle a, l_1, l_2 \rangle \in \theta_i$ or $\langle a + \pi, l_1, l_2 \rangle \in \theta_i$. Thus, the weight of the arc in the depicting $\theta_i$ graph is equal either to the relative alignment of the edges, or to this relative alignment incremented by $\pi$, satisfying condition (5) in the definition of depicted alignment graphs.

In the third subcase, if $\langle e_1, e_2 \rangle \in \text{parallel}$, then axiom 5.3 of $T_{\Delta}$ gives $\langle l_1, l_2 \rangle \in \text{parallel}_i$. By the definition of $\theta_i$ graphs, the weight of the directed arc $(l_1, l_2)$ must be 0, and since the depicted alignment graph is a subgraph of a $\theta_i$ graph, it must have the same weight for this arc, condition (2) the definition of depicted alignment graphs is satisfied.
Now consider the case where the lines \( l_1, l_2 \) depict the same edge in the surface, so that

\[
\langle l_1, e \rangle, \langle l_2, e \rangle \in \Delta
\]

\[
\langle e, s \rangle \in \text{part}
\]

Axiom 5.4 of \( T_\Delta \) gives

\[
\langle \pi, l_1, l_2 \rangle \in \theta
\]

so that the weight of the arc in the depicting \( \theta_i \) graph is equal to \( \pi \), satisfying condition (3) in the definition of depicted alignment graphs.

\( \Box \)

We are now ready to prove the Axiomatizability Theorem for \( T_\Delta \).

**Theorem 5.9** Let \( \mathcal{M} \) be a structure in \( L_\Delta \) with the following properties:

- \( \mathcal{M} \) contains a substructure isomorphic to a structure in \( M_\text{polygon}^i \);
- \( \mathcal{M} \) contains a substructure isomorphic to a structure in \( M_\text{angle}^i \);
- \( \mathcal{M} \) is a model of \( T_{\text{scene}} \cup T_{\text{kernel}} \cup T_{fg} \cup T_\Delta \);

then \( \mathcal{M} \) is isomorphic to a structure in \( M^\Delta \).

**Proof:** We will show that any model \( \mathcal{M} \) of \( T_{\text{scene}} \cup T_{\text{kernel}} \cup T_{fg} \cup T_\Delta \) that has the properties in the hypothesis of the theorem satisfies the conditions in the definition of \( M^\Delta \).

Conditions 1, 2, and 3 in the definition of \( M^\Delta \) follow trivially from the hypothesis of the theorem. \( \mathcal{M} \) contains a unique depicted alignment graph for each depicted surface in the model, by Lemma 5.7.

Next we need to show that \( \mathcal{M} \) contains a unique depicted figure/ground graph for each depicted surface in the model that has multiple depicted edges. For every surface which has two depicted edges, Corollary 5.1 shows that for every depicted inside graph there exists a isomorphic depicting between graph, so that for every surface there exists a depicted figure/ground graph.

Theorems 5.5 and 5.6 show that the depicted figure/ground graph is unique. Further, for every surface which has two depicted edges, Theorem 5.7 shows that for every depicting between graph there exists a unique isomorphic depicted inside graph, showing that any model of \( T_{\text{scene}} \cup T_{\text{kernel}} \cup T_{fg} \cup T_\Delta \) satisfies condition (4) in the definition of \( M^\Delta \).

For every surface which has only one depicted edge, Lemma 5.6 shows that there exists a solitary depicting graph for the surface. Theorem 5.8 shows that this graph is unique. Thus, \( \mathcal{M} \) satisfies condition (5). \( \Box \)

### 5.7 Conclusions

We have presented a set of axioms for the depiction of surfaces, and shown that these axioms preserve the unique figure ground assignment of a surface. In terms of images, this property is closely related to the notion of ambiguity in an image. If we have already grouped the lines in an image by asserting that they depict edges in the same surface, then the only ambiguity over the depiction of surfaces is due to uncertainty in the figure ground relation of the depicted edges; for example, in Figure 5.4(a) and (b) if we do not know whether \( \text{inside}(e_1, e_2) \) or \( \text{outside}(e_1, e_2) \), then we cannot assign surface depiction
literals. We will later discover that this ambiguity plays a key role in the intractability of finding a model of the CardWorld theories. This property of the depiction axioms also lays the foundation for the grouping assumptions that we present later in the thesis.
Chapter 6

Additional Theories for Scenes and Images

In this chapter, we present two additional theories for scenes and depiction, as well as introducing a model-theoretic specification of the class of images which we will be considering for the remainder of the thesis. Both of the new theories play a supporting role in the development of a theory of occlusion and a characterization of assumptions about errors in edge detection. Our first task will therefore be to specify the conditions that make occlusion possible. The primary intuition that will motivate the axioms presented in this section is that a surface occludes another scene object because the occluding surface constitutes a figure. That is, one surface can occlude another surface only at positions interior to the surface. In this way, the characterization theorem for figure and ground can be extended to help define the conditions for occlusion. These intuitions will lead to the development of a theory about the interior of a surface, and the relation between the interior of a surface and any holes which the surface may have.

The second task will be a characterization of the class of images. In particular, we will be restricting our attention to images which are isomorphic to planar graphs. This will allow us to reason about the existence of models for a broad class of images. This restriction will also allow us to reason about the existence of regions within an image, without being trapped by considering particular images.

The third task will combine these two developments into a theory of depicted surfaces, which is the subject of the last section of this chapter.

6.1 Interior of a Surface

In this first section, we introduce $T_{\text{interior}}$, the CardWorld module which axiomatizes intuitions about positions being interior to a surface. The primary motivation for this theory is that we can consider the interior of a surface as the generalization of figure and ground from edges to arbitrary positions. Similarly, since surfaces may have holes, we can also generalize figure/ground assignments to positions which are interior to some hole in the surface. Both of these ideas lead to the following intuitions:
Intuition 9 The interior of the surface lies on a unique “side” of an edge in the surface. Positions on the “figure” side of an edge are interior to the surface, and positions which are on the “ground” side of every edge are not interior to the surface.

Intuition 10 Every surface has a unique set of positions which are interior to the surface, and possibly disjoint sets of positions which are interior to any holes which the surface may have.

Intuition 11 The edges which are the boundary for a hole in a surface cannot be outer edges. Positions which are on the “ground” side of a hole edge are interior to the hole of the surface.

These three intuitions form the central elements which will be formalized in the class of structures $M_{\text{interior}}$, and axiomatized in the CardWorld module $T_{\text{interior}}$.

6.1.1 Some Geometry of Euclidean Planes

Before embarking on the structures and axioms of the theory of surface interiors, we will introduce a set of relations for surfaces which are concerned with embedding a scene (i.e., a set of surfaces) into a Euclidean plane. As with earlier geometric relations, we will not be explicitly axiomatizing these relations within this thesis; we will be using the class of structures that capture the intended interpretations of the relations in Hilbert’s axiomatization of geometry.

The language $L_{\text{plane}}$ which we will adopt for surfaces is:

$$L_{\text{plane}} = \{\text{position}(p, t), \text{sweep}(e_1, e_2, t), \text{noncrossing}(s, t_1, t_2)\}$$

- The binary predicate $\text{position}(p, t)$ denotes a functional relation in which the point $p$ of a surface has a position denoted by the point $t$ in the Euclidean plane. We will informally refer to this Euclidean point as a position. It should be noted that a position can be interior to multiple surfaces, depending on the position of points in the surfaces.

- The intended interpretation of the ternary predicate $\text{sweep}(e_1, e_2, t)$ is related to the definition of relative alignment. Recall that $\theta$ was defined as the angle between two edges as the segments corresponding to the edges rotated clockwise on the pivot position which is their intersection. In both cases, we can think of two edges as dividing the plane into two subplanes – one swept out between the edges $e$ and $e'$ in the clockwise direction and the other swept out between $e'$ and $e$ in the clockwise direction. Thus, given a position $t$, we have $\text{sweep}(e, e', t)$ iff the position $t$ is in the subplane that is swept out between the edges in the clockwise direction, using the intersection point of their extensions as the pivot. For example, in Figure 6.1(b), the positions $t_1, t_2, t_3, t_4$ are all within the sweep of the edge $e_1$ to the edge $e_3$. The positions $t_2, t_3$ are within the sweep of the edge $e_4$ to the edge $e_6$, and the positions $t_1, t_2, t_4$ are within the sweep of the edge $e_2$ to the edge $e_1$.

- The intended interpretation of the ternary predicate $\text{noncrossing}(s, t_1, t_2)$ is that any segment that is drawn between the position $t_1$ and position $t_2$ within the plane cannot intersect the position
of any point in an edge of the surface \( s \). (This segment exists within the model of Hilbert's axiomatization of geometry.) It generalizes the intuition of contiguous from edges to positions. For example, in Figure 6.1(b), the positions \( t_1 \) and \( t_4 \) are noncrossing, but the positions \( t_1 \) and \( t_2 \) are not, and neither are the positions \( t_4 \) and \( t_3 \).

We can formally specify the intended interpretations of these relations using Hilbert's axiomatization of geometry.

**Definition 6.1** Let \( \mathcal{H} \) be the model of Hilbert's axiomatization of geometry, and suppose \( \mathcal{M} \) is a structure in \( \mathcal{L}_{\text{plane}} \). We will say that \( \mathcal{M} \) is in the class of structures \( \mathcal{M}_{i}^{\text{plane}} \) iff:

- \( \langle p, t \rangle \in \text{position} \iff t \text{ is a point in } \mathcal{H} \).
- If the lines containing the segments \( h(e_1) \) and \( h(e_2) \) intersect, then \( \langle e_1, e_2, t \rangle \in \text{sweep} \iff t \text{ is in the interior of the angle formed by the lines in } \mathcal{H} \) containing the segments \( h(l_1) \) and \( h(l_2) \).
- If the lines containing the segments \( h(e_1) \) and \( h(e_2) \) are parallel, then \( \langle e_1, e_2, t \rangle \in \text{sweep} \iff \text{there exists a segment containing a point } x_1 \text{ in } h(e_1), \text{ a point } x_2 \text{ in } h(e_2) \text{ such that } (x_1, h(t), x_2) \in \text{ordered} \).
- \( \langle s, t_1, t_2 \rangle \in \text{noncrossing} \iff \text{there does not exist a point } p \text{ such that } \langle p, s \rangle \in \text{part and } \langle t_1, h(p), t_2 \rangle \in \text{ordered} \)

Examples of these relations can be found using Figure 6.1. In Figure 6.1(b), there is a structure \( \mathcal{M} \in \mathcal{M}_{i}^{\text{plane}} \) such that

\[
\langle e_1, e_3, t_1 \rangle, \langle e_1, e_3, t_2 \rangle, \langle e_1, e_3, t_3 \rangle, \langle e_1, e_3, t_4 \rangle, \\
\langle e_2, e_1, t_1 \rangle, \langle e_2, e_1, t_2 \rangle, \langle e_2, e_1, t_4 \rangle, \\
\langle e_6, e_5, t_2 \rangle \in \text{sweep}
\]

Also within this structure, we have

\[\langle s, t_1, t_4 \rangle \in \text{noncrossing}\]

Note that none of the other positions illustrated in Figure 6.1(b) are noncrossing, since any segment connecting these other positions will intersect an edge of the surface.

**6.1.2 Defining the Interior of a Surface**

The nonlogical lexicon for \( T_{\text{interior}} \) is:

\[
\mathcal{L}_{\text{interior}} = \mathcal{L}_{f_3} \cup \mathcal{L}_{\text{plane}} \cup \{\text{interior}(s, t), \text{boundary}(e, t), \text{hole,interior}(s, t), \text{hole,boundary}(e, t)\}
\]

- The binary predicate \( \text{interior}(s, t) \) denotes the relation defined over a surface and the set of positions in the plane. Intuitively, the position \( t \) is on the "figure" side of the edges of the surface \( s \).
Figure 6.1: Surfaces used to illustrate the intended interpretations of predicates in $L_{\text{plane}}$ and $L_{\text{interior}}$. 
• The binary predicate $\text{boundary}(e, t)$ denotes the relation defined over pairs of edges in a surface and the set of positions in the plane. Intuitively, an edge is a boundary edge for a position iff the position lies on the "figure" side of the edge and there exist points in the edge that is noncrossing with respect to the position.

• The binary predicate $\text{hole.interior}(s, t)$ denotes the relation defined over a surface and the set of positions in the plane. Intuitively, the position $t$ is on the "ground" side of the hole edges of the surface $s$.

• The binary predicate $\text{hole.boundary}(e, t)$ denotes the relation defined over pairs of edges in a surface and the set of positions in the plane. Intuitively, this relation is the dual of the boundary relation - an edge is a hole boundary edge for a position iff the position lies on the "ground" side of the edge in the hole.

Consider the surface in Figure 6.1(b). The positions $t_1$ and $t_4$ are interior to the surface, while the position $t_2$ is $\text{hole.interior}$ to the surface. Edges $e_1, e_2, e_4, e_5$ are the boundary edges for position $t_4$, while edges $e_1, e_2, e_3, e_6$ are the boundary edges for position $t_1$. All three hole edges ($e_4, e_5, e_6$) are hole boundary edges for the position $t_2$. The position $t_3$ is neither interior nor $\text{hole.interior}$ to the surface.

We now need to specify the various structures which are isomorphic to the extensions of these new relations. Notice that all of them are relations between positions and scene objects (either edges or surfaces). Thus, the structures we need will contain sets of positions:

**Definition 6.2** Given a structure $\mathcal{M}$ in $\mathcal{L}_{\text{interior}}$ which is an extension of a structure in $\mathcal{M}^g_1$, a patch is a maximal set $\Sigma$ of positions such that for any two positions $t_1, t_2 \in \Sigma$, we have

$\langle s, t_1, t_2 \rangle \in \text{noncrossing}$

We will denote the set of all patches in a structure by $\mathcal{T}$.

Consider the surfaces in Figure 6.1. For the L-shape in (a), there are two patches which will be of particular interest; these are shown in Figure 6.2(a) and (b). Intuitively, all positions in the shaded portion of the surface in Figure 6.2(a) have the same boundary edges - $e_1, e_2, e_5, e_6$. All positions in the shaded portion of the surface in Figure 6.2(b) have the same boundary edges - $e_3, e_4, e_5, e_6$. The set of positions which are interior to the surface is the union of the positions which are in either of these shaded portions.

For the triangle shape with a hole, there are four patches which will be of particular interest; these are shown in Figure 6.2(c), (d), (e), (f). The set of positions in the shaded portion in Figure 6.2(f) are all interior to the hole in the surface; the hole boundary edges for these positions are $e_4, e_5, e_6$.

There are two major kinds of structures which we will use to characterize the extensions of the relations in $\mathcal{L}_{\text{interior}}$ - edge-position graphs and hole-position graphs. Following Intuitions 9 and 11, these graphs must capture the relationship between patches and the figure/ground relations (inside and
Figure 6.2: Examples of the interior of a surface.
outside) for a surface. We therefore must first characterize substructures which capture the relevant properties of the figure/ground assignment for the surface.

**Definition 6.3** Given a structure $M$ in $M_1^{f9}$, the contiguous-inside subgraph for a surface $s$ is the subgraph of the inside graph for $s$ consisting of all edges which are contiguous with each other.

**Definition 6.4** The contiguous-outside subgraph for a surface $s$ is the subgraph of the outside graph for $s$ consisting of all hole edges which are contiguous with each other.

The contiguous-inside graph for the surface in Figure 6.1(a) is shown in Figure 6.3(a). The contiguous-inside graph for the surface in Figure 6.1(b) is shown in Figure 6.3(b). The contiguous-outside graph for the surface in Figure 6.1(b) is shown in Figure 6.3(c).

Given these two classes of substructures for figure/ground assignments, we can define the edge-position graph as specifying a mapping between edges of a surface and patches (sets of positions). In the following definition of edge-position graphs, the first two conditions specify the relationship between the edges of the edge-position graph and the figure/ground assignments for these edges. The last two conditions characterize the sets of positions which can possibly be in a patch that is an element of an edge-position graph. The patches in an edge-position graph are not arbitrary sets of positions in the plane: these patches are defined with respect to two contiguous edges, the subplane which is the sweep of these two edges, and the condition that every position in the patch be noncrossing with respect to the position of a point in one of the edges.
**Definition 6.5** Given a structure $\mathcal{M}$ in $\mathcal{L}_{\text{interior}}$ which is an extension of a structure in $\mathcal{M}_{i}^{g}$ and which is an extension of a structure in $\mathcal{M}_{i}^{\text{plane}}$, an edge-position graph for a surface is a graph $G = (N, A)$ with the following properties:

1. $G$ is a directed bipartite graph, with $N \subseteq \mathcal{E} \cup \mathcal{T}$, \(^1\) and if $(x, y) \in A$ then $x \in \mathcal{E}, y \in \mathcal{T}$.

2. The set of edges in a clique of the contiguous-inside graph for the surface is equivalent to the set of edges connected to the same patch in the edge-position graph.

3. The set of edges connected to the same patch in the edge-position graph forms a clique in the contiguous-inside graph of the surface.

4. If $\Pi$ is a patch which is a node in $G$ and $(e, \Pi) \in A$, then for every point $p$ in $e$ with position $t_1$ and every position $t \in \Pi$, we have

   $$(s, t, t_1) \in \text{noncrossing}$$

5. For any two edges $e_1, e_2$ connected to the same patch $\Pi$, and for any position $t \in \Pi$, we have

   $$(e_1, e_2, t) \in \text{sweep}$$

The edge-position graph for the surface in Figure 6.2(a) is shown in Figure 6.4. In this graph, there are two patches $(\Pi_1, \Pi_2)$. Consider the contiguous-inside and contiguous-outside graphs from Figure 6.3. The maximal cliques in the graph in Figure 6.3(a) are:

$$\{e_1, e_2, e_3, e_4\}, \{e_3, e_4, e_5, e_6\}$$

The first set of edges is connected to the patch $\Pi_1$ in the edge-position graph, and the second set of edges is connected to the patch $\Pi_2$.

The edge-position graph for the surface in Figure 6.2(b) is shown in Figure 6.5(a). In this graph, there are three patches $(\Pi_1, \Pi_2, \Pi_3)$. Consider the contiguous-inside and contiguous-outside graphs from Figure 6.3. The maximal cliques in the graph in Figure 6.3(b) are:

$$\{e_1, e_2, e_3, e_4\}, \{e_1, e_2, e_3, e_4\}, \{e_1, e_2, e_3, e_5\}$$

The first set of edges is connected to the patch $\Pi_3$ in the edge-position graph, and the second set of edges is connected to the patch $\Pi_1$, and the third set of edges is connected to the patch $\Pi_2$.

Analogously, the hole-position graph specifies a mapping between edges of a surface and patches (sets of positions). However, whereas the definition of edge-position graphs used the contiguous-inside graph to identify the "figure" side of two edges, the definition of hole-position graph uses the contiguous-outside graph to identify the "ground" side of two hole edges in the surface. In the following definition of hole-position graphs, the first two conditions specify the relationship between the edges of the hole-position graph and the figure/ground assignments for these edges. The last two conditions characterize the sets of positions which can possibly be in a patch that is an element of a hole-position graph. Again, the patches in a hole-position graph are not arbitrary sets of positions in the plane; these patches are

\(^1\)Recall that $\mathcal{E}$ is the set of edges in a structure $\mathcal{M}$.
Figure 6.4: Edge-position graph for the surface in Figure 6.2(a).

Figure 6.5: Edge-position and hole-position graphs for the surface in Figure 6.2(b).
defined with respect to two contiguous edges, the subplane which is the sweep of these two edges, and the condition that every position in the patch be noncrossing with respect to the position of a point in one of the edges.

Definition 6.6 Given a structure \( \mathcal{M} \) in \( \mathcal{L}_{\text{interior}} \) which is an extension of a structure in \( \mathcal{M}_i^3 \) and which is an extension of a structure in \( \mathcal{M}_i^{\text{plane}} \), a hole-position graph for a surface is a graph \( G = (N, A) \) with the following properties:

1. \( G \) is a directed bipartite graph, with \( N \subseteq (\mathcal{E} - \mathcal{E}_{\text{outer}}) \cup \mathcal{T} \), and if \( (x, y) \in A \) then \( x \in \mathcal{E} - \mathcal{E}_{\text{outer}} \) and \( y \in \mathcal{T} \).

2. the set of edges in a clique of the contiguous-outside graph for the surface is equivalent to the set of edges connected to the same patch in the hole-position graph.

3. The set of edges connected to the same patch in the hole-position graph forms a clique in the contiguous-outside graph of the surface.

4. If \( \Pi \) is a patch which is a node in \( G \) and \( (e, \Pi) \in A \), then for every point \( p \) in \( e \) with position \( t_1 \) and every position \( t \in \Pi \), we have

\[ \langle s, t, t_1 \rangle \in \text{noncrossing} \]

5. For any two edges \( e_1, e_2 \) connected to the same patch \( \Pi \), and for any position \( t \in \Pi \), we have

\[ \langle e_1, e_2, t \rangle \in \text{sweep} \]

The hole-position graph for the surface in Figure 6.2(b) is shown in Figure 6.5(b). In this graph, there is one patch \( (\Pi_4) \). There is a unique maximal clique in the contiguous-outside graph in Figure 6.3(c):

\[ \{e_4, e_5, e_6\} \]

This set of edges is connected to the patch \( \Pi_4 \) in the hole-position graph.

In addition to specifying the interior of a single surface, we will also need to capture the relationship among multiple surfaces in a scene, in which surfaces overlap and occlude each other. In all of the previous CardWorld theories we have considered, all scene relations have been restricted to scene objects within the same surface. For example, the figure/ground relations inside and outside are restricted to edges which are part of the same surface. Similarly, the classes of structures which we have defined so far in this chapter consider surfaces in isolation e.g. the edge-position graph of a particular surface. However, since scenes contain surfaces which can possibly overlap each other, there will be positions which are interior to multiple surfaces. We therefore need to characterize the extensions of interior and \( \text{hole.interior} \) with respect to the edge-position and hole-position graphs for multiple surfaces.

For example, consider the overlapping surfaces in Figure 6.6(a) \(^3\). The position \( t_1 \) is interior to both surfaces \( s_1 \) and \( s_2 \), while the position \( t_2 \) is interior only to surface \( s_1 \). It should therefore be the case that there exists a boundary edge for \( t_1 \) which is part of \( s_2 \) and which contains a point whose position is also interior to \( s_1 \).

---

\(^2\)Recall that \( \mathcal{E} \) is the set of edges in a structure \( \mathcal{M} \), and that \( \mathcal{E}_{\text{outer}} \) is the set of outer edges in a structure \( \mathcal{M} \).

\(^3\)Note that these are representations of a scene, in which all edges of both surfaces are displayed; they are not images.
Intuition 12: If two surfaces have overlapping interiors, then there exists a point in a boundary edge of one surface which is also interior to the other surface.

Similarly, consider the overlapping surfaces in Figure 6.6(b), in which we have two surfaces with holes. Position $t_2$ is interior to surface $s_2$ but interior to the hole of surface $s_1$. Intuitively, there should be a point in a boundary edge for $t_2$ which shares this property:

Intuition 13: If the interior of one surface overlaps with the hole of another surface, then there exists a point in a boundary edge of the surface which is also interior to a hole of the other surface.

Finally, in Figure 6.6(b), position $t_1$ is interior to the holes of both surfaces, while $t_2$ is interior to the surface $s_2$. In this case, there should be point in a hole boundary edge for $t_1$ which is also interior to the hole of $s_1$:

Intuition 14: If two surfaces have overlapping hole interiors, then there exists a point in a hole boundary edge of one surface which is also interior to the hole of the other surface.

We now have enough structures to define $M_i^{\text{interior}}$, which will be the class of intended models for the CardWorld module $T^{\text{interior}}$. Conditions (6) - (10) in the definition of $M_i^{\text{interior}}$ specify the extensions of the relations denoted by the predicates in $L^{\text{interior}}$. Conditions (3) - (5) specify the relationship between surfaces in a structure and its associated edge-position and hole-position graphs.

Intuitions 12-14 are formalized by conditions (10) - (12) in the definition of $M_i^{\text{interior}}$. They are essentially constraints on how the edge-position and hole-position graphs for multiple surfaces in a scene can be combined within a structure. Note that the sets of edges in the edge-position and hole-position graphs for distinct surfaces must be disjoint (since an edge is part of a unique surface); however, there may exist patches which are elements of different edge-position and hole-position graphs. It is the relationship among these patches which needs to be captured and formalized.
Definition 6.7 Let $\mathcal{M}_{1}^{\text{interior}}$ be a set of structures in $\mathcal{L}_{\text{interior}}$ with the following properties: For each structure $\mathcal{M} \in \mathcal{M}_{1}^{\text{interior}}$,

1. $\mathcal{M}$ is an extension of a structure in $\mathcal{M}^{fg}$.
2. $\mathcal{M}$ is an extension of a structure in $\mathcal{M}_{\text{plane}}$.
3. Each surface has a unique edge-position graph.
4. A surface has $n$ disconnected sets of non-outer edges iff the surface is associated with $n$ disjoint hole-position graphs.
5. The set of patches in the edge-position graph of a surface are disjoint from the set of patches in all hole-position graphs of the surface.
6. $(e, t) \in \text{boundary}$

iff the position $t$ is contained in a patch which is connected to the edge $e$ in an edge-position graph in $\mathcal{M}$.
7. $(e, t) \in \text{hole\_boundary}$

iff the position $t$ is contained in a patch which is connected to the edge $e$ in a hole-position graph in $\mathcal{M}$.
8. $(s, t) \in \text{interior}$

iff the position $t$ is contained in a patch which is a node of the edge-position graph of the surface $s$.
9. $(s, t) \in \text{hole\_interior}$

iff the position $t$ is contained in a patch which is a node of a hole-position graph of the surface $s$.
10. If an edge in a surface $s_1$ is connected to a patch $\Pi_1$ in the edge-position graph, and there exists a patch $\Pi_2$ which is an element of the edge-position graph of a distinct surface $s_2$ such that $\Pi_1 \cap \Pi_2 \neq \emptyset$, then there exists a point $p$ which is part of a boundary edge of $s_2$ and whose position is also contained in $\Pi_1$.
11. If an edge in a surface $s_1$ is connected to a patch $\Pi_1$ in the edge-position graph, and there exists a patch $\Pi_2$ which is an element of a hole-position graph of a distinct surface $s_2$ such that $\Pi_1 \cap \Pi_2 \neq \emptyset$, then there exists a point $p$ which is part of a boundary edge of $s_1$ and whose position is also contained in $\Pi_2$.
12. If an edge in a surface $s_1$ is connected to a patch $\Pi_1$ in a hole-position graph, and there exists a patch $\Pi_2$ which is an element of a hole-position graph of a distinct surface $s_2$ such that $\Pi_1 \cap \Pi_2 \neq \emptyset$, then there exists a point $p$ which is part of a hole boundary edge of $s_1$ and whose position is also contained in $\Pi_2$.

In the next section, we propose an axiomatization of this class of structures using $\mathcal{L}_{\text{interior}}$.

Theorem 6.1 Existence Theorem for $\mathcal{M}_{1}^{\text{interior}}$

The class of structures $\mathcal{M}_{1}^{\text{interior}}$ exists and is nonempty.
**Proof:** Since structures in $\mathcal{M}_{\text{interior}}$ are the union of edge-position graphs and hole-position graphs, we will first need to show that these classes of graphs exist.

Recall that edge-position graphs are bipartite graphs such that the edges in the neighborhood of a patch are also edges in a clique of the contiguous-inside graph, which is a subgraph of the $\text{inside}$ graph for the surface.

By the Existence Theorem for $\mathcal{M}_{\text{interior}}$, we know that $\text{inside}$ graphs exist. Let $\chi$ be the set of cliques in some $\text{inside}$ graph $G_1$. For any clique, we can construct a bipartite graph $G_2 = (X, Y, A)$ such that $X$ is the set of edges in the surface and the edges in the clique are connected to the same element of $X$. Thus, the conditions are satisfied, and edge-position graphs exist.

Similarly, hole-position graphs are bipartite graphs such that the edges in the neighborhood of a patch are also edges in a clique of the contiguous-outside graph, which is a subgraph of the $\text{outside}$ graph for the surface.

By the Existence Theorem for $\mathcal{M}_{\text{interior}}$, we know that $\text{outside}$ graphs exist. Let $\chi$ be the set of cliques in some $\text{outside}$ graph $G_1$. For any clique, we can construct a bipartite graph $G_2 = (X, Y, A)$ such that $X$ is the set of edges in the surface and the edges in the clique are connected to the same element of $X$. Thus, the conditions are satisfied, and hole-position graphs exist.

Finally, conditions (10)-(12) specify how the edge-position and hole-position graphs for distinct surfaces can be combined within a structure in $\mathcal{M}_{\text{interior}}$, that is, only substructures which satisfy these conditions can be combined. However, since we have already shown that the substructures themselves exist, their union must also exist. ☐

### 6.1.3 Axiomatization of $T_{\text{interior}}$

The axioms of $T_{\text{interior}}$ are presented in Figure 6.8.

The definition of boundary relates the predicates in $L_{\text{plane}}$ with the figure/ground relations. We need to represent the notion that the interior of a surface corresponds to the figure, or inside of the surface, and the exterior corresponds to the ground. The intuition that two edges forming the boundary of the surface relative to some position $t$ must be figure corresponds to the first two conjuncts in the definition. The third conjunct specifies the intuition about position – there cannot be any edges that partition the position $t$ from either of the boundary edges. In this way, we can consider the definition of boundary to be a generalization of the notion of figure and ground from edges to positions.

Given the definition of boundary edges, the definition of interior is simply equivalent to the existence of a boundary edge. Thus, the interior corresponds to the figure and the exterior (those positions not interior) corresponds to ground.

The definition of hole-boundary also relates the predicates in $L_{\text{plane}}$ with the figure/ground relations. In this case, we need to represent the notion that the interior of a hole in the surface corresponds to the "ground", or outside of the surface. The intuition that two hole edges forming the boundary of the hole relative to some position $t$ must be "ground" corresponds to the first part of the definition. Note that only non-outer (i.e. hole) edges can possibly be boundary edges for a hole. The last conjunct specifies the intuition about position – there cannot be any edges that partition the position $t$ from either of the boundary edges. In this way, we can consider the definition of boundary to be a generalization of the
An edge is a boundary edge for a position iff the position lies on the "figure" side of the edge and there exists a point in the edge that is noncrossing with respect to the position.

\[(\forall e, t) \text{boundary}(e, t) \equiv (\exists e_2) \text{contiguous}(e, e_2) \land (\text{inside}(e, e_2) \equiv \text{sweep}(e, e_2)) \land ((\forall s, p, t') \text{part}(p, e) \land \text{position}(p, t') \land \text{part}(e, s) \supset \text{noncrossing}(s, t, t'))\]  
(6.1)

A position is interior to a surface iff there exist edges in the surface that are boundary edges for the position.

\[(\forall s, t) \text{interior}(s, t) \equiv (\exists e) \text{part}(e, s) \land \text{boundary}(e, t)\]  
(6.2)

A hole edge is a hole boundary edge for a position iff the position lies on the "ground" side of the edge and there exists a point in the edge that is noncrossing with respect to the position.

\[(\forall e, t) \text{hole-boundary}(e, t) \equiv \]

\[(\exists e_2) \text{contiguous}(e, e_2) \land \neg \text{outer}(e) \land \neg \text{outer}(e_2) \land (\text{outside}(e, e_2) \equiv \text{sweep}(e, e_2)) \land ((\forall s, p, t') \text{part}(p, e) \land \text{position}(p, t') \land \text{part}(e, s) \supset \text{noncrossing}(s, t, t'))\]  
(6.3)

A position is interior to a hole of a surface iff there exist edges in the surface that are hole boundary edges for the position.

\[(\forall s, t) \text{hole-interior}(s, t) \equiv (\exists e) \text{part}(e, s) \land \text{hole-boundary}(e, t)\]  
(6.4)

Figure 6.7: Conservative definitions for $T_{\text{interior}}$. 
Boundary edges and hole boundary edges are disjoint sets.

\[(\forall e, t) \text{boundary}(e, t) \cup \neg\text{hole\_boundary}(e, t)\] (6.5)

Hole boundary edges for the same position must be connected.

\[(\forall e_1, e_2, t) \text{hole\_boundary}(e_1, t) \land \text{hole\_boundary}(e_2, t) \cup \text{connected}(e_1, e_2)\] (6.6)

If two surfaces have overlapping interiors, then there exists a point in a boundary edge of one surface which is also interior to the other surface.

\[(\forall s_1, s_2, t_1, t_2) \text{interior}(s_1, t_1) \land \text{interior}(s_2, t_1) \land \neg\text{interior}(s_2, t_2) \land \text{interior}(s_1, t_2) \cup\]

\[(\exists e, p, t_3) \text{boundary}(e, t_1) \land \text{part}(e, s_2) \land \text{part}(p, e) \land \text{position}(p, t_3) \land \text{interior}(s_1, t_3)\] (6.7)

If the interior of one surface overlaps with the hole of another surface, then there exists a point in a boundary edge of the surface which is also interior to a hole of the other surface.

\[(\forall s_1, s_2, t_1, t_2) \text{interior}(s_1, t_1) \land \text{hole\_interior}(s_2, t_1) \land \neg\text{interior}(s_1, t_2) \land \text{hole\_interior}(s_2, t_2) \cup\]

\[(\exists e, p, t_3) \text{boundary}(e, t_1) \land \text{part}(e, s_2) \land \text{part}(p, e) \land \text{position}(p, t_3) \land \text{hole\_interior}(s_2, t_3)\] (6.8)

If two surfaces have overlapping hole interiors, then there exists a point in a boundary edge of one surface which is also interior to the hole of the other surface.

\[(\forall s_1, s_2, t_1, t_2) \text{interior}(s_1, t_1) \land \text{hole\_interior}(s_2, t_1) \land \text{hole\_interior}(s_1, t_2) \land \text{hole\_interior}(s_2, t_2) \cup\]

\[(\exists e, p, t_3) \text{hole\_boundary}(e, t_2) \land \text{part}(e, s_1) \land \text{part}(p, e) \land \text{position}(p, t_3) \land \text{hole\_interior}(s_2, t_3)\] (6.9)

Figure 6.8: The axioms of \(T_{\text{interior}}\).
notion of figure and ground from edges to positions. Analogously to the \textit{interior} relation, the definition of \textit{hole\_interior} is simply equivalent to the existence of a hole boundary edge.

We also want to ensure that positions interior to a surface are not interior to a hole of the surface. Axiom 6.5 ensure this, since no edge can be both a boundary edge and a hole boundary edge. Similarly, we want to ensure that the positions interior to one hole of a surface are not interior to other holes of the surface. Axiom 6.6 ensures this, since all hole boundary edges of a position must be connected, while each hole of a surface intuitively corresponds to a different disjoint set of hole edges.

Axioms 6.7 - 6.9 capture the constraints among the positions which are \textit{interior} and/or \textit{hole\_interior} to multiple surfaces. They are a straightforward axiomatization of conditions (10)-(12) in the definition of $\mathcal{M}_i^\text{interior}$.

\textbf{Example:} Consider the surface in Figure 6.1(b). There exists a model $\mathcal{M}$ of $T_{\text{interior}}$ such that

\begin{equation*}
\{e_1, e_2, e_3, e_4, e_5, e_6, s_1, t_1, t_2, t_3, t_4\} \subseteq M \quad \text{and which has the following properties:}
\end{equation*}

Since

\begin{equation*}
\langle e_2, e_1, t_1 \rangle \in \text{sweep}, \langle e_1, e_2 \rangle \in \text{contiguous}, \langle e_2, e_1 \rangle \in \text{inside}
\end{equation*}

and all positions in either edge are noncrossing with respect to $t_1$, by the definition of boundary edges we have

\begin{equation*}
\langle e_1, t_1 \rangle, \langle e_2, t_1 \rangle \in \text{boundary}
\end{equation*}

and by the definition of \textit{interior} we have

\begin{equation*}
\langle s_1, t_1 \rangle \in \text{interior}
\end{equation*}

Since

\begin{equation*}
\langle e_4, e_6, t_2 \rangle \in \text{sweep}, \langle e_4, e_6 \rangle \in \text{contiguous}, \langle e_4 \rangle \not\in \text{outer}, \langle e_6 \rangle \not\in \text{outer}, \langle e_4, e_6 \rangle \in \text{outside}
\end{equation*}

and all positions in either edge are noncrossing with respect to $t_2$, by the definition of hole boundary edges we have

\begin{equation*}
\langle e_4, t_2 \rangle, \langle e_6, t_2 \rangle \in \text{hole\_boundary}
\end{equation*}

and by the definition of \textit{interior} we have

\begin{equation*}
\langle s_1, t_2 \rangle \in \text{hole\_interior}
\end{equation*}

Note that we have

\begin{equation*}
\langle s_1, t_2 \rangle \not\in \text{interior}
\end{equation*}

since the only edges which are both contiguous and inside are the outer edges $e_1, e_2, e_3$, and for any of these edges, a segment connecting the position $t_2$ must cross another edge in the surface, violating the last conjunct in the definition of \textit{interior}. 
Similarly, we have

\( (s_1, t_3) \notin \text{interior} \)

since there are no edges which are both inside and which contain the position in their sweep. □

### 6.1.4 Satisfiability of \( T_{\text{interior}} \)

We now show that the relations in \( L_{\text{interior}} \) have the extensions specified in the definition of \( M_{\text{interior}} \).

Essentially, we need to show that the definition of boundary edges is satisfied by edge-position graphs, and that the definition of hole boundary edges is satisfied by hole-position graphs.

**Theorem 6.2** Any structure in \( M_{\text{interior}} \) is a model of \( T_{\text{interior}} \cup T_{fg} \cup T_{\text{scene}} \).

**Proof:** We first consider the relation boundary.

Let \( \mathcal{M} \in M_{\text{interior}} \) and suppose for any variable assignment.

\[
\mathcal{M}, \sigma \models \text{boundary}(e, t)
\]

By definition of \( M_{\text{interior}} \), the position \( \sigma(t) \) is contained in a patch which is connected to \( \sigma(e) \) in an edge-position graph \( G \) in \( \mathcal{M} \), which exists by condition (3).

For the converse direction, suppose that

\[
\mathcal{M}, \sigma \models (\exists e_2 \) contiguous\((e, e_2) \land (\text{inside}(e, e_2) \equiv \text{sweep}(e, e_2, t)) \land (\exists s, p, t') \) part\((p, e) \land \text{position}(p, t') \land \text{part}(e, s) \land \text{noncrossing}(s, t, t')
\]

By definition of \( M_{\text{interior}} \), this defines an edge-position graph \( G = (N, A) \) for the surface \( \sigma(s) \) such that

\[
(\sigma(e), \sigma(t)) \in A
\]

Therefore, \( \sigma(t) \) is contained in a patch connected to \( \sigma(e) \) in \( G \), and hence

\[
\mathcal{M}, \sigma \models \text{boundary}(e, t)
\]

Now consider the definition of interior.

Let \( \mathcal{M} \in M_{\text{interior}} \) and suppose for any variable assignment,

\[
\mathcal{M}, \sigma \models \text{interior}(s, t)
\]

By definition of \( M_{\text{interior}} \), the position \( \sigma(t) \) is contained in a patch which is a node in the edge-position graph for \( s \). By the definition of edge-position graphs, this patch must be connected to some edge \( e \in M \) such that \( (e, \sigma(s)) \in \text{part} \). The definition of boundary then entails \( (e, \sigma(t)) \in \text{boundary} \), so that

\[
\mathcal{M}, \sigma \models (\exists e) \text{part}(e, s) \land \text{boundary}(e, t)
\]

For the converse direction, suppose

\[
\mathcal{M}, \sigma \models (\exists e) \text{part}(e, s) \land \text{boundary}(e, t)
\]

This is equivalent to the existence of an edge \( e \in M \) such that

\[
(e, \sigma(s)) \in \text{part}, (e, \sigma(t)) \in \text{boundary}
\]
By the definition of boundary and the definition of $M_i^{\text{interior}}$, $\sigma(t)$ is contained in a patch which is connected to $e$ in the edge-position for $\sigma(s)$, so consequently

$$M, \sigma \models \text{interior}(s, t)$$

Now consider the definition of hole_boundary.

Let $M \in M_i^{\text{interior}}$ and suppose for any variable assignment,

$$M, \sigma \models \text{hole_boundary}(e, t)$$

By definition of $M_i^{\text{interior}}$, the position $\sigma(t)$ is contained in a patch which is connected to $\sigma(e)$ in a hole-position graph $G$ in $M$, which exists by condition (4).

For the converse direction, suppose that

$$M, \sigma \models \neg \text{boundary}(e, t)$$

and

$$M, \sigma \models \neg \text{hole_boundary}(e, t)$$

By definition of $M_i^{\text{interior}}$, this defines a hole-position graph $G = (N, A)$ for the surface $\sigma(s)$ such that

$$(\sigma(e), \sigma(t)) \in A$$

Therefore, $\sigma(t)$ is contained in a patch connected to $\sigma(e)$ in $G$, and hence

$$M, \sigma \models \text{hole_boundary}(e, t)$$

Now consider the definition of hole_interior.

Let $M \in M_i^{\text{interior}}$ and suppose for any variable assignment,

$$M, \sigma \models \text{hole_interior}(s, t)$$

By definition of $M_i^{\text{interior}}$, the position $\sigma(t)$ is contained in a patch which is a node in a hole-position graph for $s$. By the definition of hole-position graphs, this patch must be connected to some non-outer edge $e \in M$ such that $(e, \sigma(s)) \in \text{part}$. The definition of hole_boundary then entails $(e, \sigma(t)) \in \text{hole_boundary}$, so that

$$M, \sigma \models (\exists e) \text{part}(e, s) \wedge \text{hole_boundary}(e, t)$$

For the converse direction, suppose

$$M, \sigma \models (\exists e) \text{part}(e, s) \wedge \text{hole_boundary}(e, t)$$

This is equivalent to the existence of a non-outer edge $e \in M$ such that

$$(e, \sigma(s)) \in \text{part}, (e, \sigma(t)) \in \text{hole_boundary}$$

By the definition of hole_boundary and the definition of $M_i^{\text{interior}}$, $\sigma(t)$ is contained in a patch which is connected to $e$ in a hole-position for $\sigma(s)$, so consequently

$$M, \sigma \models \text{hole_interior}(s, t)$$

Consider axiom 6.5:

$$M \models (\forall e, t) \text{boundary}(e, t) \supset \neg \text{hole_boundary}(e_2, t)$$

iff for any variable assignment $\sigma$ there does not exist an edge $\sigma(e)$ and a position $\sigma(t)$ such that

$$(\sigma(e), \sigma(t)) \in \text{boundary}, (\sigma(e), \sigma(t)) \in \text{hole_boundary}$$
This follows from condition (5) in the definition of $\mathcal{M}^\text{interior}_t$, which requires that the set of boundary edges for a position is disjoint from the set of hole boundary edges.

Consider axiom 6.6:

$$\mathcal{M} \models (\forall e_1, e_2, t) \text{ hole\_boundary}(e_1, t) \land \text{ hole\_boundary}(e_2, t) \supset \text{ connected}(e_1, e_2)$$

iff for any variable assignment $\sigma$ there do not exist edges $\sigma(e_1), \sigma(e_2)$ and a position $\sigma(t)$ such that

$$\langle \sigma(e_1), \sigma(t) \rangle, \langle \sigma(e_2), \sigma(t) \rangle \in \text{ hole\_boundary}$$

$$\langle \sigma(e_1), \sigma(e_2) \rangle \notin \text{ connected}$$

This follows from condition (4) in the definition of $\mathcal{M}^\text{interior}_t$, which requires that the sets of hole boundary edges for a position are disjoint iff there are disconnected sets of non-outer edges.

Finally, consider the axioms of $T^\text{interior}$, which specify the relationships among multiple surfaces in a scene.

Axiom 6.7:

$$\mathcal{M} \models (\forall s_1, s_2, t_1, t_2) \text{ interior}(s_1, t_1) \land \text{ interior}(s_2, t_1) \land \neg \text{ interior}(s_2, t_2) \land \text{ interior}(s_2, t_1) \supset$$

$$(\exists e, p, t_3) \text{ boundary}(e, t_1) \land \text{ part}(e, s_2) \land \text{ part}(p, e) \land \text{ position}(p, t_3) \land \text{ interior}(s_1, t_3)$$

is equivalent to the following: if for any surfaces $\sigma(s_1), \sigma(s_2) \in M$ and any positions $\sigma(t_1), \sigma(t_2) \in M$, if

$$\mathcal{M}, \sigma \models \text{ interior}(s_1, t_1) \land \text{ interior}(s_2, t_1) \land \neg \text{ interior}(s_2, t_2) \land \text{ interior}(s_2, t_1)$$

then there exists an edge $\sigma(e)$, a point $\sigma(p)$, and a position $\sigma(t_3)$ such that

$$\mathcal{M}, \sigma \models \text{ boundary}(e, t_1) \land \text{ part}(e, s_2) \land \text{ part}(p, e) \land \text{ position}(p, t_3) \land \text{ interior}(s_1, t_3)$$

The antecedent is satisfied by $\mathcal{M}$ iff the position $\sigma(t_1)$ is both interior to $\sigma(s_1)$ and interior to $\sigma(s_2)$ and the position $\sigma(t_2)$ is interior to $\sigma(s_2)$ but not interior to $\sigma(s_1)$. By the definition of interior, this is satisfied iff $\sigma(s_1)$ and $\sigma(s_2)$ are connected to patches that have a nonempty intersection in their respective edge-position graphs.

The consequent is satisfied iff the position $\sigma(t_3)$ of the point $\sigma(p)$ is also interior to $\sigma(s_1)$, which is satisfied iff $\sigma(t_3)$ is contained in the same patch containing $\sigma(t_2)$. This holds by condition (10) in the definition of $\mathcal{M}^\text{interior}_t$, so that axiom 6.7 is satisfied by $\mathcal{M}$.

Axiom 6.8:

$$\mathcal{M} \models (\forall s_1, s_2, t_1, t_2) \text{ hole\_interior}(s_1, t_1) \land \text{ hole\_interior}(s_2, t_1) \land \neg \text{ hole\_interior}(s_2, t_2) \land \text{ hole\_interior}(s_2, t_1) \supset$$

$$(\exists e, p, t_3) \text{ boundary}(e, t_1) \land \text{ part}(e, s_2) \land \text{ part}(p, e) \land \text{ position}(p, t_3) \land \text{ hole\_interior}(s_2, t_3)$$

is equivalent to the following: if for any surfaces $\sigma(s_1), \sigma(s_2) \in M$ and any positions $\sigma(t_1), \sigma(t_2) \in M$, if

$$\mathcal{M}, \sigma \models \text{ hole\_interior}(s_1, t_1) \land \text{ hole\_interior}(s_2, t_1) \land \text{ hole\_interior}(s_2, t_2) \land \neg \text{ hole\_interior}(s_2, t_1)$$

then there exists an edge $\sigma(e)$, a point $\sigma(p)$, and a position $\sigma(t_3)$ such that

$$\mathcal{M}, \sigma \models \text{ boundary}(e, t_1) \land \text{ part}(e, s_2) \land \text{ part}(p, e) \land \text{ position}(p, t_3) \land \text{ hole\_interior}(s_1, t_3)$$

The antecedent is satisfied by $\mathcal{M}$ iff the position $\sigma(t_1)$ is both interior to $\sigma(s_1)$ and hole\_interior to $\sigma(s_2)$ and the position $\sigma(t_2)$ is hole\_interior to $\sigma(s_2)$ but not interior to $\sigma(s_1)$. By the definitions of interior and hole\_interior, this is satisfied iff $\sigma(s_1)$ is connected to a patch in its edge-position graph, and this patch has a nonempty intersection with a patch that is connected to $\sigma(s_2)$ its hole-position graph.
The consequent is satisfied if the position $\sigma(t_3)$ of the point $\sigma(p)$ is also $\text{hole\_interior}$ to $\sigma(s_1)$, which is satisfied if $\sigma(t_2)$ is contained in the same patch containing $\sigma(t_2)$. This holds by condition (11) in the definition of $M_i^\text{interior}$, so that axiom 6.7 is satisfied by $\mathcal{M}$.

Axiom 6.9:

$$\mathcal{M} \models (\forall s_1, s_2, t_1, t_2) \text{interior}(s_1, t_1) \land \text{hole\_interior}(s_2, t_1) \land \text{hole\_interior}(s_2, t_2) \land \text{hole\_interior}(s_1, t_2) \supset$$

$$(\exists e, p, t_3) \text{ hole\_boundary}(e, t_1) \land \text{part}(e, s_2) \land \text{part}(p, e) \land \text{position}(p, t_3) \land \text{hole\_interior}(s_2, t_3)$$

is equivalent to the following: if for any surfaces $\sigma(s_1), \sigma(s_2) \in \mathcal{M}$ and any positions $\sigma(t_1), \sigma(t_2) \in \mathcal{M}$, if

$$\mathcal{M}, \sigma \models \text{interior}(s_1, t_1) \land \text{hole\_interior}(s_2, t_1) \land \text{hole\_interior}(s_2, t_2) \land \text{hole\_interior}(s_1, t_2)$$

then there exists an edge $\sigma(e)$, a point $\sigma(p)$, and a position $\sigma(t_3)$ such that

$$\mathcal{M}, \sigma \models \text{hole\_boundary}(e, t_2) \land \text{part}(e, s_2) \land \text{part}(p, e) \land \text{position}(p, t_3) \land \text{hole\_interior}(s_2, t_3)$$

The antecedent is satisfied by $\mathcal{M}$ iff the position $\sigma(t_1)$ is both $\text{interior}$ to $\sigma(s_1)$ and $\text{hole\_interior}$ to $\sigma(s_2)$ and and the position $\sigma(t_2)$ is $\text{hole\_interior}$ to $\sigma(s_1)$ and $\text{hole\_interior}$ to $\sigma(s_2)$. By the definitions of $\text{interior}$ and $\text{hole\_interior}$, this is satisfied if $\sigma(s_1)$ is connected to a patch in its hole-position graph, and this patch has a nonempty intersection with a patch that is connected to $\sigma(s_2)$ its hole-position graph.

The consequent is satisfied iff the position $\sigma(t_2)$ of the point $\sigma(p)$ is also $\text{hole\_interior}$ to $\sigma(s_1)$, which is satisfied iff $\sigma(t_2)$ is contained in the same patch containing $\sigma(t_2)$. This holds by condition (12) in the definition of $M_i^\text{interior}$, so that axiom 6.7 is satisfied by $\mathcal{M}$. □

### 6.1.5 Characterization Theorems for $T_{\text{interior}}$

Using the characterization theorem for figure and ground from the previous chapter, we can show that the interior and exterior of a surface are well-defined. In the following theorem, we show that every surface has a unique assignment of $\text{interior}$ to all positions, thus generalizing the notion of figure and ground from edges to positions. In this sense, the assignment of $\text{interior}$ for a surface follows from the unique figure ground assignment given to every surface. If there is no ambiguity with respect to the position of points in a surface then there will be no ambiguity with respect to the extension of $\text{interior}$.

**Theorem 6.3** If $\mathcal{M}, \mathcal{M}'$ are models of $T_{\text{scene}} \cup T_{fg} \cup T_{\text{interior}}$ that have the same domain, agree on the extensions of the relations in $L_{\text{plane}}$, and agree on the extensions of the relations in $L_{\text{polygon}}$, then they agree on the extensions of $\text{interior}, \text{boundary}, \text{hole\_interior}, \text{and hole\_boundary}$.

**Proof:** Since the models agree on the extensions of relations in $L_{\text{plane}},$ they agree on the extension of $\text{contiguous}$. By the Characterization Theorem for Figure and Ground, $\mathcal{M}, \mathcal{M}'$ agree on the extensions of $\text{inside}$ and $\text{outside}$. Thus, the $\text{contiguous}$ edges in $\mathcal{M}$ must have the same figure ground assignment in $\mathcal{M}'$.

Since $\mathcal{M}, \mathcal{M}'$ agree on the extensions of all relations in $L_{\text{plane}},$ for any variable assignment $\sigma$ we have

$$\mathcal{M}, \sigma \models \text{sweep}(e, e', t_1) \land \text{noncrossing}(s, t_1, t_2)$$

iff

$$\mathcal{M}', \sigma \models \text{sweep}(e, e', t_1) \land \text{noncrossing}(s, t_1, t_2)$$

Since $\mathcal{M}, \mathcal{M}'$ agree on the extensions of all relations in $L_{\text{plane}},$ for any variable assignment $\sigma$ we have
Thus, we must have
\[ M, \sigma \models \text{boundary}(e, t) \]
iff
\[ M', \sigma \models \text{boundary}(e, t) \]
From the definition of interior, it is easy to see that this also implies
\[ M, \sigma \models \text{interior}(s, t) \]
iff
\[ M', \sigma \models \text{interior}(s, t) \]
A similar argument holds for hole\_interior and hole\_boundary. □

This result is a constraint on the entire set of structures in \( M_\text{interior} \), specifying the conditions under which structures have isomorphic extensions of relations denoted by the predicates in \( L_\text{interior} \): it corresponds to Intuition 10.

We will need the following lemma to show that the edges connected to the same patch in an edge-position graph form a clique in the contiguous-inside graph for the surface.

**Lemma 6.1** Let \( M \) be a model of \( T\text{scene} \cup T_f \cup T_\text{interior} \) and which is also an extension of a structure in \( M_\text{plane} \); then we have
\[
M \models (\forall e_1, e_2, t) \ \text{boundary}(e_1, t) \land \text{contiguous}(e_1, e_2) \land (\text{inside}(e_1, e_2) \equiv \text{sweep}(e, e_2, t))
\]
\[
\land ((\forall s, p, t') \ \text{part}(p, e_2) \land \text{position}(p, t') \land \text{part}(e_2, s) \supset \text{noncrossing}(s, t, t'))
\]
\[
\supset \text{boundary}(e_2, t)
\]

Thus, we can use the edges in the contiguous-inside subgraph to determine the set of boundary edges for a position.

**Proof:** Suppose for some variable assignment \( \sigma \),
\[ M, \sigma \models \text{boundary}(e_1, t) \]
By the definition of boundary, then there exists another edge \( \sigma(e_2) \) such that
\[ M, \sigma \models \text{contiguous}(e_1, e_2) \land (\text{inside}(e_1, e_2) \equiv ((\exists a) \ \theta(a, e_1, e_2) \land a < \pi) \]
By \( T_f \), we have
\[ M, \sigma \models \text{contiguous}(e_2, e_1) \land (\neg \text{inside}(e_2, e_1) \equiv ((\exists a) \ \theta(a, e_1, e_2) \land a < \pi) \]
In any structure in \( M_\text{plane} \), we have
\[ M \models (\forall e_1, e_2, t) \ \text{sweep}(e_1, e_2, t) \supset \neg \text{sweep}(e_2, e_1, t) \]
so that we have
\[ M, \sigma \models \text{contiguous}(e_2, e_1) \land (\neg \text{inside}(e_2, e_1) \equiv \neg \text{sweep}(e_2, e_1, t)) \]
If we assume that
\[ M, \sigma \models ((\forall s, p, t') \ \text{part}(p, e_2) \land \text{position}(p, t') \land \text{part}(e_2, s) \supset \text{noncrossing}(s, t, t')) \]
then we can see that
\[ M, \sigma \models \text{boundary}(e_2, t) \]
□
Similarly, we will need the following lemma to show that the edges connected to the same patch in an hole-position graph form a clique in the contiguous-outside graph for the surface.

**Lemma 6.2** Let $M$ be a model of $T_{\text{scene}} \cup T_{fg} \cup T_{\text{interior}}$ and which is also an extension of a structure in $M_{i}^{\text{plane}}$; then we have

$$M, \sigma \models \text{hole\_boundary}(e_1, t) \land \text{contiguous}(e_1, e_2) \land (\text{outside}(e, e_2) \equiv \text{sweep}(e, e_2, t))$$

$$\land ((\forall s, p, t') \text{ part}(p, e_2) \land \text{position}(p, t') \land \text{part}(e_2, s) \supset \text{noncrossing}(s, t, t'))$$

Thus, we can use the edges in the contiguous-outside subgraph to determine the set of hole-boundary edges for a position.

**Proof:** Suppose for some variable assignment $\sigma$,

$$M, \sigma \models \text{hole\_boundary}(e_1, t)$$

By the definition of $\text{hole\_boundary}$, then there exists another edge $\sigma(e_2)$ such that

$$M, \sigma \models \text{contiguous}(e_1, e_2) \land (\text{outside}(e_1, e_2) \equiv ((\exists a) \theta(a, e_1, e_2) \land a < \pi)$$

By $T_{fg}$, the figure/ground assignment for hole edges is the inverse of the figure/ground assignment for outer edges, so we have

$$M, \sigma \models \text{contiguous}(e_2, e_1) \land (\neg\text{outside}(e_2, e_1) \equiv ((\exists a) \theta(a, e_1, e_2) \land a < \pi)$$

In any structure in $M_{i}^{\text{plane}}$, we have

$$M \models (\forall e_1, e_2, t) \text{sweep}(e_1, e_2, t) \supset \neg\text{sweep}(e_2, e_1, t)$$

so that we have

$$M, \sigma \models \text{contiguous}(e_2, e_1) \land (\neg\text{outside}(e_2, e_1) \equiv \neg\text{sweep}(e_2, e_1, t))$$

If we assume that

$$M, \sigma \models (\forall s, p, t') \text{ part}(p, e_2) \land \text{position}(p, t') \land \text{part}(e_2, s) \supset \text{noncrossing}(s, t, t'))$$

then we can see that

$$M, \sigma \models \text{hole\_boundary}(e_2, t)$$

$\Box$

We are now ready to prove the main result of this section – the structures in $M_{i}^{\text{interior}}$ characterize the models of $T_{\text{interior}}$. Equivalently, this result shows that the class of structures in $M_{i}^{\text{interior}}$ is axiomatized by $T_{\text{interior}}$.

**Theorem 6.4** Let $M$ be a structure in $L_{\text{interior}}$ with the following properties:

- $M$ contains a substructure isomorphic to a structure in $M_{i}^{\text{polygon}}$;
- $M$ contains a substructure isomorphic to a structure in $M_{i}^{\text{plane}}$;
Then \( U \) is isomorphic to a structure in \( M_{\text{interior}} \).

**Proof:** Let \( M \) be a model of \( T_{\text{scene}} \cup T_{\text{fg}} \cup T_{\text{interior}} \) that satisfies the properties in the hypothesis of the theorem.

We will first characterize the extension of **boundary** within \( M \). Suppose that for any variable assignment \( \sigma \),

\[ M, \sigma \models \text{boundary}(e, t) \]

By the definition of **boundary** (axiom 6.1), we have

\[ M, \sigma \models (\exists e_2) \text{contiguous}(e, e_2) \land (\text{inside}(e, e_2) \equiv \text{sweep}(e, e_2, t)) \]

\[ \land ((\forall s, p, t') \text{part}(p, e) \land \text{position}(p, t') \land \text{part}(e, s) \supset \text{noncrossing}(s, t, t')) \]

By the third conjunct, we can construct a bipartite graph \( G = (N, A) \) such that \( N = E \cup T_1 \) (where \( T_1 \) is the set of positions in \( M \)) and \((\sigma(e), \sigma(t)) \in A \) iff

\[ (\langle \sigma(p), \sigma(e) \rangle \in \text{part}, (\sigma(e), \sigma(s)) \in \text{part}, (\sigma(p), \sigma(t')) \in \text{position}, (\sigma(s), \sigma(t), \sigma(t')) \in \text{noncrossing} \]

By Lemma 6.1, if

\[ (\sigma(e_2), \sigma(e)) \in \text{contiguous}, (\sigma(e_2), \sigma(e)) \in \text{inside} \]

then we must have \((\sigma(e), \sigma(t)) \in A \) iff \((\sigma(e_2), \sigma(t)) \in A \). Thus, edges in a clique of the contiguous-inside graph are connected to the same position in the graph \( G \), and edges in \( G \) which are connected to the same position are also contained in a clique of the contiguous-inside graph.

Given the graph \( G \), we can now construct a graph \( G' = (N', A') \) which is isomorphic to an edge-position graph such that

\[ (\sigma(e), \sigma(t)) \in A \Leftrightarrow (\sigma(e), \Pi) \in A' \text{ and } \sigma(t) \in \Pi \]

Thus,

\[ M, \sigma \models \text{boundary}(e, t) \]

iff the position \( \sigma(t) \) is contained in a patch which is connected to the edge \( \sigma(e) \) in an edge-position graph in \( M \), and \( M \) satisfies condition (6) in the definition of \( M_{\text{interior}} \).

From the preceding argument, it is easy to see that if

\[ M, \sigma \models (\forall s, t) \text{interior}(s, t) \equiv (\exists e) \text{part}(e, s) \land \text{boundary}(e, t) \]

then \((\sigma(s), \sigma(t)) \in \text{interior} \) iff the position \( \sigma(t) \) is contained in a patch which is a node of the edge-position graph of \( \sigma(s) \), so that \( M \) satisfies condition (8) in the definition of \( M_{\text{interior}} \).

We next characterize the extension of **hole_boundary** within \( M \). Suppose that for any variable assignment \( \sigma \),

\[ M, \sigma \models \text{hole_boundary}(e, t) \]

By the definition of **hole_boundary** (axiom 6.3), we have

\[ M, \sigma \models (\exists e_2) \text{contiguous}(e, e_2) \land (\text{outside}(e, e_2) \equiv \text{sweep}(e, e_2, t)) \land \neg \text{outer}(e) \land \neg \text{outer}(e_2) \]

\[ \land ((\forall s, p, t') \text{part}(p, e) \land \text{position}(p, t') \land \text{part}(e, s) \supset \text{noncrossing}(s, t, t')) \]

By the final conjunct in the axiom, we can construct a bipartite graph \( G = (N, A) \) such that \( N = E \cup T_1 \) (where \( T_1 \) is the set of positions in \( M \)) and \((\sigma(e), \sigma(t)) \in A \) iff

\[ (\sigma(e)) \in \text{outer} \]
CHAPTER 6. ADDITIONAL THEORIES FOR SCENES AND IMAGES

$\langle \sigma(p), \sigma(e) \rangle \in \text{part}, \langle \sigma(e), \sigma(s) \rangle \in \text{part}, \langle \sigma(p), \sigma(t') \rangle \in \text{position}, \langle \sigma(s), \sigma(t), \sigma(t') \rangle \in \text{noncrossing}$

By Lemma 6.2, if

$\langle \sigma(e_2), \sigma(e) \rangle \in \text{contiguous}, \langle \sigma(e_2), \sigma(e) \rangle \in \text{outside}$

then we must have $\langle \sigma(e), \sigma(t) \rangle \in A$ iff $\langle \sigma(e_2), \sigma(t) \rangle \in A$. Thus, edges in a clique of the contiguous-outside graph are connected to the same position in the graph $G$, and edges in $G$ which are connected to the same position are also contained in a clique of the contiguous-outside graph.

Given the graph $G$, we can now construct a graph $G' = (V', E')$ which is isomorphic to a hole-position graph such that

$\langle \sigma(e), \sigma(t) \rangle \in A \iff \langle \sigma(e), \Pi \rangle \in A'$ and $\sigma(t) \in \Pi$

Thus,$M, \sigma \models \text{hole boundary}(e, t)$

iff the position $\sigma(t)$ is contained in a patch which is connected to the edge $\sigma(e)$ in an hole-position graph in $M$, and $M$ satisfies condition (7) in the definition of $M_i \text{interior}$.

From the preceding argument, it is easy to see that if

$M, \sigma \models (\forall s, t \text{ hole interior}(s, t) \equiv (\exists e) \text{ part}(e, s) \land \text{hole boundary}(e, t))$

then $\langle \sigma(s), \sigma(t) \rangle \in \text{hole interior}$ iff the position $\sigma(t)$ is contained in a patch which is a node of the hole-position graph of $\sigma(s)$, so that $M$ satisfies condition (9) in the definition of $M_i \text{interior}$.

By Theorem 6.3, condition (3) in the definition of $M_i \text{interior}$ is satisfied by $M$, since for each surface $m$, $G$ is guaranteed a nonempty extension of $\text{interior}$.

Suppose

$T_{\text{interior}} \models (\exists s, t) \text{ interior}(s, t) \land \text{hole interior}(s, t)$

Then by the definition of $\text{interior}$ and $\text{hole interior}$, we have

$T_{\text{interior}} \models (\exists e, t) \text{ boundary}(e, t) \land \text{hole boundary}(e, t)$

which violates axiom 6.5. Thus, condition (4) in the definition of $M_i \text{interior}$ is satisfied by $M$, since for each surface $m$, $G$ is guaranteed a nonempty extension of $\text{interior}$.

It is easy to see that condition (5) in the definition of $M_i \text{interior}$ is satisfied by $M$, by axiom 6.6.

Finally, consider the axioms of $T_{\text{interior}}$ which specify the relationships among multiple surfaces in a scene.

$M, \sigma \models (\forall s_1, s_2, t_1, t_2) \text{ interior}(s_1, t_1) \land \text{interior}(s_2, t_1) \land \neg \text{interior}(s_2, t_2) \land \text{interior}(s_2, t_1) \lor$

$(\exists e, p, t_3) \text{ boundary}(e, t_1) \land \text{part}(e, s_2) \land \text{part}(p, e) \land \text{position}(p, t_3) \land \text{interior}(s_1, t_3)$

is equivalent to the following: if for any surfaces $\sigma(s_1), \sigma(s_2) \in M$ and any positions $\sigma(t_1), \sigma(t_2) \in M$, if

$M, \sigma \models \text{interior}(s_1, t_1) \land \text{interior}(s_2, t_1) \land \neg \text{interior}(s_2, t_2) \land \text{interior}(s_2, t_1)$

then there exists an edge $\sigma(e)$, a point $\sigma(p)$, and a position $\sigma(t_3)$ such that

$M, \sigma \models \text{boundary}(e, t_1) \land \text{part}(e, s_2) \land \text{part}(p, e) \land \text{position}(p, t_3) \land \text{interior}(s_1, t_3)$

The antecedent is satisfied by $M$ iff the position $\sigma(t_1)$ is both $\text{interior}$ to $\sigma(s_1)$ and $\text{interior}$ to $\sigma(s_2)$ and the position $\sigma(t_2)$ is $\text{interior}$ to $\sigma(s_2)$ but not $\text{interior}$ to $\sigma(s_1)$. By the definition of $\text{interior}$, this is satisfied iff $\sigma(s_1)$ and $\sigma(s_2)$ are connected to patches in their respective edge-position graphs which have a nonempty intersection (which contains in the position $\sigma(t_2)$).

The consequent is satisfied iff the position $\sigma(t_3)$ of the point $\sigma(p)$ is also $\text{interior}$ to $\sigma(s_1)$, which is satisfied iff $\sigma(t_3)$ is contained in the same patch containing $\sigma(t_2)$. 
Thus, condition (10) in the definition of $M_i^{\text{interior}}$ is satisfied by $M$.

Next, suppose

$$M, \sigma \models (\forall s_1, s_2, t_1, t_2) \text{interior}(s_1, t_1) \land \text{hole.interior}(s_2, t_1) \land \text{hole.interior}(s_2, t_2) \land \neg \text{interior}(s_2, t_1) \supset$$

$$\exists e, p, t_3 \text{ boundary}(e, t_1) \land \text{part}(e, s_2) \land \text{part}(p, e) \land \text{position}(p, t_3) \land \text{interior}(s_2, t_3)$$

This is equivalent to the following: if for any surfaces $\sigma(s_1), \sigma(s_2) \in M$ and any positions $\sigma(t_1), \sigma(t_2) \in M$, if

$$M, \sigma \models \text{interior}(s_1, t_1) \land \text{hole.interior}(s_2, t_1) \land \text{hole.interior}(s_2, t_2) \land \neg \text{interior}(s_2, t_1)$$

then there exists an edge $\sigma(e)$, a point $\sigma(p)$, and a position $\sigma(t_3)$ such that

$$M, \sigma \models \text{boundary}(e, t_1) \land \text{part}(e, s_2) \land \text{part}(p, e) \land \text{position}(p, t_3) \land \text{hole.interior}(s_2, t_3)$$

The antecedent is satisfied by $M$ iff the position $\sigma(t_1)$ is both $\text{interior}$ to $\sigma(s_1)$ and $\text{hole.interior}$ to $\sigma(s_2)$ and and the position $\sigma(t_2)$ is $\text{hole.interior}$ to $\sigma(s_2)$ but not $\text{interior}$ to $\sigma(s_1)$. By the definitions of $\text{interior}$ and $\text{hole.interior}$, this is satisfied iff $\sigma(s_1)$ is connected to a patch in its edge-position graph, and this patch has a nonempty intersection with a patch that is connected to $\sigma(s_2)$ its hole-position graph.

The consequent is satisfied iff the position $\sigma(t_3)$ of the point $\sigma(p)$ is also $\text{hole.interior}$ to $\sigma(s_1)$, which is satisfied iff $\sigma(t_3)$ is contained in the same patch containing $\sigma(t_2)$. Thus, condition (11) in the definition of $M_i^{\text{interior}}$ is satisfied by $M$.

Finally, suppose

$$M, \sigma \models (\forall s_1, s_2, t_1, t_2) \text{interior}(s_1, t_1) \land \text{hole.interior}(s_2, t_1) \land \text{hole.interior}(s_2, t_2) \land \text{hole.interior}(s_1, t_2) \supset$$

$$\exists e, p, t_3 \text{ hole.boundary}(e, t_1) \land \text{part}(e, s_2) \land \text{part}(p, e) \land \text{position}(p, t_3) \land \text{hole.interior}(s_2, t_3)$$

This is equivalent to the following: if for any surfaces $\sigma(s_1), \sigma(s_2) \in M$ and any positions $\sigma(t_1), \sigma(t_2) \in M$, if

$$M, \sigma \models \text{interior}(s_1, t_1) \land \text{hole.interior}(s_2, t_1) \land \text{hole.interior}(s_2, t_2) \land \text{hole.interior}(s_1, t_2)$$

then there exists an edge $\sigma(e)$, a point $\sigma(p)$, and a position $\sigma(t_3)$ such that

$$M, \sigma \models \text{hole.boundary}(e, t_2) \land \text{part}(e, s_2) \land \text{part}(p, e) \land \text{position}(p, t_3) \land \text{hole.interior}(s_2, t_3)$$

The antecedent is satisfied by $M$ iff the position $\sigma(t_1)$ is both $\text{interior}$ to $\sigma(s_1)$ and $\text{hole.interior}$ to $\sigma(s_2)$ and and the position $\sigma(t_2)$ is $\text{hole.interior}$ to $\sigma(s_2)$ and $\text{hole.interior}$ to $\sigma(s_1)$. By the definitions of $\text{interior}$ and $\text{hole.interior}$, this is satisfied iff $\sigma(s_1)$ is connected to a patch in its hole-position graph, and this patch has a nonempty intersection with a patch that is connected to $\sigma(s_2)$ its hole-position graph.

The consequent is satisfied iff the position $\sigma(t_3)$ of the point $\sigma(p)$ is also $\text{hole.interior}$ to $\sigma(s_1)$, which is satisfied iff $\sigma(t_3)$ is contained in the same patch containing $\sigma(t_2)$. Thus, condition (12) in the definition of $M_i^{\text{interior}}$ is satisfied by $M$.

Since all of the conditions in the definition of $M_i^{\text{interior}}$ are satisfied by $M$, we have $M \in M_i^{\text{interior}}$. \(\square\)
6.2 Images as Planar Graphs

In order to prove the theorems characterizing the relationships between the various occlusion and depiction theories and assumptions, we will first need to define the conditions necessary for the existence of regions. In other words, we need a set of conditions that all images must satisfy: in this section we introduce and discuss these conditions on images.

It should be realized that these conditions are not used in constructing models of an image, since we assume closure over image predicates and objects. However, they are necessary to prove properties that hold for images in general (or at least some class of images), given a set of occlusion and depiction assumptions. For example, we will later need to prove that regions are guaranteed to exist if we assume that there are no errors in edge detection: we cannot do this unless we have axioms that define what regions are. An alternative way to think about these image conditions is that they define the class of images which we will consider throughout this work. In particular, we will be restricting ourselves to planar images (i.e. images that are isomorphic to the class of planar graphs).

In earlier chapters, we have axiomatized each class of structures using some CardWorld module. We will be taking a different approach with the treatment of images. Rather than proving the soundness and completeness of an axiomatization of a class of structures for images, we will only be providing a model-theoretic specification of this class. One of the primary reasons for this manoeuvre is that the class of planar graphs is not first-order definable, and the second-order sentence which does define this class is rather complicated. 4

The nonlogical lexicon \( L_{\text{planar}} \) for planar images will be:

- The binary predicate \( \text{inclusion}(r,t) \) denoting the relation which is satisfied by all positions which are in the interior of a region.

- The binary predicate \( \text{exclusion}(r,t) \) denoting the relation which is satisfied by all positions which are not in the interior of a region.

- The binary predicate \( \text{border}(l,r) \) denoting the relation with the intended interpretation that \( l \) is a line in the boundary, or border, of the region \( r \). This reflects the intuition that a region has a closed boundary consisting of intersecting lines.

These predicates will be given an intended interpretation using the notions of planar graphs and subdivisions. The following definition is taken from [Berg et al. 97]:

**Definition 6.8** A planar subdivision is a partitioning of a Euclidean plane which is induced by the embedding of a planar graph.

A face is a maximal connected subset of the plane that doesn't contain a point on an arc or node of the planar graph; its boundary is formed by the lines of the planar graph.

---

4In \( T_fg \) we encountered the second-order definition for the connected relation; since this sentence was rather straightforward, we chose to explicitly include the axiom in \( T_fg \).
If a node in the graph is the endpoint of an arc in the graph, then we say that the node and the arc are incident. Similarly, a face and an arc on its boundary are incident, and a face and a node if its boundary are incident.

Note that a face is a connected subset of the plane, just as patches are connected sets of positions in the plane. This correspondence will play a key role in the next section.

We can now define the intended class of images, which will be those that are isomorphic to planar graphs:

**Definition 6.9** A planar image is one which satisfies the following properties:

- Lines in an image are isomorphic to the arcs in a planar graph $G$.
- For lines which intersect two more lines at distinct junctions in the image, there exist two distinct nodes in $G$.
- For lines which do not intersect any line, there exist two nodes in the graph connected by an arc; this arc is disconnected from any other nodes in $G$.
- For lines which intersect a unique line in the image, there exists one node in $G$ corresponding to the junction in the image, and another node in $G$ which is connected only to the junction in the image.

For example, consider the image in Figure 6.10(b). The planar graph corresponding to this image is shown in Figure 6.9.

The class of structures $M_{\text{region}}^{\text{planar}}$ will specify the extensions of the relations in $L_{\text{planar}}$ with respect to the properties of planar graphs and subdivisions. In particular, conditions (9) and (10) in the definition specify the relationship between the incidence relation for planar subdivisions and the incidence relation in for image objects.

**Definition 6.10** Let $M_{\text{region}}^{\text{planar}}$ be the class of structures in $L_{\text{planar}} \cup L_{\text{kernel}}$ such that for any structure $M \in M_{\text{region}}^{\text{planar}}$, the following conditions holds:

1. $M$ is an extension of a structure in $M_{\text{kernel}}$.
2. All images in $M$ are planar images.
3. For every region \( r \in M \) there exists a unique face in a planar subdivision.

4. For every face in a planar subdivision, there exists a region \( r \in M \).

5. \( \langle l, r \rangle \in \text{border} \) iff the line \( l \) is incident with the face associated with region \( r \). (Thus, lines in cycles of the planar graph of an image are the border lines for some region in the image.)

6. \( \langle r, t \rangle \in \text{inclusion} \) iff the position \( t \) is contained in the face associated with the region \( r \).

7. \( \langle r, t \rangle \in \text{exclusion} \) iff the position \( t \) is not contained in the face associated with the region \( r \).

8. Pixels which are in a border line of a region are not included in the region.

9. Border lines are contained in their regions, i.e.
   \( \text{if } \langle l, r \rangle \in \text{border}, \text{ then } \langle l, r \rangle \in \text{in} \).

10. All pixels which are contained in a region are either contained in a border line or their positions are included in the region, i.e., if \( \langle q, r \rangle \in \text{in} \), then either there exists a line \( l \) such that
   \[ \langle q, l \rangle \in \text{in}, \langle l, r \rangle \in \text{in} \]
   or there is a position \( t \) such that
   \[ \langle q, t \rangle \in \text{position}, \langle r, t \rangle \in \text{inclusion} \]

**Example:** Consider the image in Figure 6.10(a). There exists a structure \( M \) in \( L_{\text{planar}} \) with the following properties:

Each of the regions \( r_1, r_2, r_3, r_4 \) correspond to a planar subdivision.

The border lines of \( r_1 \) are:
\[ l_1, l_2, l_3, l_4, l_5, l_6, l_7 \]

The border lines of \( r_2 \) are:
\[ l_2, l_3, l_{13}, l_{12}, l_{17}, l_{10}, l_9, l_8 \]

The border lines of \( r_3 \) are:
\[ l_5, l_{15}, l_{16}, l_{11}, l_{17}, l_{12}, l_{14} \]

The border lines of \( r_4 \) are:
\[ l_4, l_{13}, l_{14} \]

Note that in each case, the border lines for a region form a cycle. Also note that for each border line is in the region corresponding to the planar subdivision.

Finally, note that
\[ \langle r_1, t_1 \rangle, \langle r_2, t_2 \rangle, \langle r_3, t_3 \rangle \in \text{inclusion} \]
\[ \langle r_1, t_2 \rangle, \langle r_2, t_3 \rangle, \langle r_3, t_1 \rangle \in \text{exclusion} \]

Using this class of images, we can now characterize how the interior of a depicted surface corresponds to the planar subdivision which is included in the region that depicts the surface.
Figure 6.10: Examples of regions in planar images.
6.3 Depicting the Interior of a Surface

The structures in $\mathcal{M}_{\text{planar}}$ define regions solely in terms of image relations. However, regions are also depicted surfaces, so that properties of regions should intuitively be related to depicted scene properties of surfaces. In particular, it would be nice if we could use a set of axioms analogous to the depicted figure/ground axioms we used in $T_\Delta$. However, a major difference between the depiction of regions and the axioms of $T_\Delta$ results from the fact that lines can be included in a region, yet not be border lines for that region. For example, in Figure 6.10(b) we have

$$I \models \text{in}(l_4, r_1) \land \neg \text{border}(l_4, r_1)$$

$$\land (\forall q, t) \text{in}(q, l_4) \land \text{position}(q, t) \supset \text{inclusion}(r_1, t)$$

This is particularly problematic since all edges form the boundary of a surface, yet not all lines need depict boundary edges of the depicted surface.

In this section we consider an axiomatization of depicted surfaces which includes the relationship between the scene relation interior for surfaces, and the image relation inclusion for regions. These axioms constitute the CardWorld module $T_{\text{interior}}$, and they will play a critical role in Chapter 8, when we provide a complete characterization of assumptions related to errors in edge detection.

Since we will be axiomatizing the depiction of the interior of a surface by planar images, the language $L_{\text{region}}$ which we will be considering in this section will be the union of the languages for the scene and image theories we have considered so far:

$$L_{\text{region}} = L_{\text{interior}} \cup L_{\text{planar}}$$

6.3.1 Intuitions for Regions as Depicted Surfaces

Since we are considering how relations in $T_{\text{interior}}$ correspond to relations in $L_{\text{planar}}$, we can begin by considering the positions of points and the positions of the corresponding pixels which depict these points:

**Intuition 15** Depiction of a scene preserves the position of depicted points.

Thus, we are considering images which are orthographic projections of the scene. Of course, we could generalize to other possible projections by modifying this condition.

**Intuition 16** Surfaces can be depicted even when all of their parts are not depicted. However, in such a case, there must be positions in the depicting region which are interior to some depicted surface.

The following example illustrates this intuition:

**Example:** Let $I$ be the image in Figure 6.11(c). There is structure $\mathcal{M} \in L_{\text{region}}$ such that $\mathcal{M} \models I$ and

$$<r_1, s_1>, <r_3, s_1> \in \Delta$$
\( (s_1, t_1) \in \text{interior}, (r_3, t_1) \in \text{inclusion} \)

Intuitively, it should not be possible for \( r_3 \) to be background and not depict a surface. In the case of Figure 6.11(c), \( r_3 \) can either depict \( s_1 \) or some other surface that is occluding \( s_1 \). \( \square \)

**Intuition 17** Inclusion in a depicting region is not equivalent to being interior to the depicted surface. They will only be equivalent if every line in the border of the region depicts a boundary edge and no point in the surface has a position included in the region.

The following examples indicate that the depiction relation between interior and inclusion is not one of equivalence: in particular, there are positions interior to some surface that are not included in the region that depicts the surface.

**Example:** Let \( I \) be the image in Figure 6.11(a) where

\[
I \models \text{inclusion}(r_1, t_1) \wedge \text{inclusion}(r_1, t_2) \wedge \text{inclusion}(r_1, t_3)
\]

There is a structure \( M \in \mathcal{L}_{\text{region}} \) which is a model of \( T_{\text{kernel}} \cup T_{\phi} \cup T_{\Delta} \cup T_{\text{interior}} \cup I \) such that

\[
(r_1, s_1), (r_1, s_2) \in \Delta
\]
\[
(s_1, t_1), (s_2, t_2) \in \text{interior}
\]
\[
(p_1, s_1) \in \text{part}, (p_1, t_3) \in \text{position}
\]
\[
(s_1, t_2), (s_2, t_1) \notin \text{interior}
\]

In this structure, there exist positions interior to the region depicting the surface \( s_1 \), yet which are not interior to \( s_1 \), and similarly for the surface \( s_2 \). In particular, note that the point \( p_1 \) in \( s_1 \) has a position which is included in the region \( r_1 \), rather than being in the border line of the region. \( \square \)

**Example:** Let \( I \) be the image Figure 6.11(d) where

\[
I \models \text{inclusion}(r_2, t_1) \wedge \text{border}(l_1, r_1) \wedge \text{border}(l_2, r_1)
\]

There is a structure \( M \in \mathcal{L}_{\text{region}} \) such that \( M \models I \) and

\[
(r_1, s_1), (r_2, s_2) \in \Delta
\]
\[
(l_1, e_1), (l_2, e_2) \in \Delta
\]
\[
(e_1, s_2), (e_2, s_2) \in \text{part}
\]
\[
(s_1, t_1) \in \text{interior}
\]

In this example, the position \( t_1 \) is interior to the surface \( s_1 \), yet it is not included in the region \( r_1 \) that depicts \( s_1 \). Notice that in this case, there are border lines \( l_1, l_2 \) of \( r_1 \) which do not depict edges in \( s_1 \). Rather, these lines depict edges in a surface which intuitively is occluding \( s_1 \). \( \square \)
Figure 6.11: Examples for intuitions about the depiction of the interior of surfaces.
Example: Let $I$ be the image Figure 6.11(b) where

$$I \models \text{inclusion}(r_1, t_1) \land \text{inclusion}(r_2, t_2)$$

There is a structure $\mathcal{M} \in \mathcal{L}_{\text{region}}$ such that $\mathcal{M} \models I$ and

$$\langle r_1, s_1 \rangle, \langle r_2, s_1 \rangle \in \Delta$$

yet there does not exist any edge depicted by $I_1$. In this structure, we have

$$\langle s_1, t_1 \rangle, \langle s_1, t_2 \rangle \in \text{interior}$$

Note that in this model, $I_1$ is an error in edge detection, since it does not depict an edge in the scene. □

**Intuition 18** Regions can also depict the holes of a surface, but such regions usually do not depict the surface. A position interior to a hole can be included in the depicting region only if there is a border line of the region which does not depict a hole boundary edge, or there is a hole edge containing a point whose position is included in the depicting region.

Even with isolated surfaces, regions do not always depict surfaces. For instance, if the surface has a hole in it, as in Figure 6.12(a), then the hole will also be “depicted” by a region, even though the hole is not part of the figure but rather part of the background. Of course, such a region should not depict the surface itself. However, there are examples where this is the case.

Example: Let $I$ be the image in Figure 6.12(b). There exists a structure $\mathcal{M} \in \mathcal{L}_{\text{region}}$ such that $\mathcal{M} \models I$ and $\{p_1, e_1, s_1\} \subseteq M$ such that

$$\langle r_1, s_1 \rangle \in \Delta, \langle l_1, e_1 \rangle \in \Delta, \langle e_1, s_1 \rangle \in \text{part}$$

$$\langle s_1, t_1 \rangle \in \text{hole.interior}, \langle r_1, t_1 \rangle \in \text{inclusion},$$

$$\langle e_1 \rangle \in \text{outer}, \langle p_1, s_1 \rangle \in \text{part}$$

and there does not exist any pixel $q$ such that $\langle q, p_1 \rangle \in \Delta$.

In this structure, there exists a position which is interior to the hole of the surface, yet which is also included in the region depicting the surface. However, this is due to the fact that a point in a hole edge is not depicted, so that there is a “gap” in the boundary of the region which should correspond to the hole. □

We next define the class of structures which will formalize these intuitions.
Figure 6.12: Examples for intuitions about the depiction of the interior of holes in surfaces.
6.3.2 Structures for Regions and Depicted Surfaces

In this section, we present the classes of structures for capturing the intuitions about the depiction of surfaces with respect to their interiors and the interiors of any hole which the surfaces may have.

We are primarily concerned with the relationship between the extension of the scene relations interior and hole-interior and the image relation inclusion. In terms of structures, this is equivalent to considering the relationship between the edge-position and hole-position graphs of a depicted surface, and the planar subdivisions for regions depicting the surface. As we did in the definition of \( \mathcal{M}_i^{\text{interior}} \), the structures we are considering are therefore substructures of existing scene or image structures. In particular, we consider the substructures of the planar subdivisions for regions depicting surfaces, and then combine these with the structures for surfaces from \( \mathcal{M}_i^{\text{interior}} \). This leads to the following three definitions:

**Definition 6.11** A depicting planar subgraph for a surface \( s \) is the subgraph of the planar image graph in \( \mathcal{M} \) such that there is a 1-1 correspondence between the set of faces of the planar image graph and the set of regions which depict \( s \).

**Definition 6.12** A depicted interior graph is the union of the edge-position graph for a depicted surface and the depicting planar subgraph for the surface.

**Definition 6.13** A depicted hole graph is the union of the hole-position graphs for the surface and the depicting planar subgraph for the surface.

No new predicate symbols are being added to the language of the structures in \( \mathcal{M}_i^{\text{region}} \). All image relations are characterized in \( \mathcal{M}_i^{\text{planar}} \), and all scene relations are characterized in \( \mathcal{M}_i^{\text{interior}} \). As with \( \mathcal{M}_i^{\Delta} \), the classes of structures we are considering are essentially a restriction of the structures in \( \mathcal{M}_i^{\text{kernel}} \). In particular, we are placing restrictions on the depiction of surfaces by regions.

**Definition 6.14** Let \( \mathcal{M} \) be a structure in \( \mathcal{L}_\text{plane} \). Two sets \( P, Q \) of elements in \( \mathcal{M} \) are position-isomorphic if for every element \( p \in P \) and \( q \in Q \),

\[
(q, p) \in \Delta \implies (q, t) \in \text{position iff } (p, t) \in \text{position}.
\]

This is a formalization of the intuition that depiction of a scene preserves the position of depicted points.

The intuitions for depicting the interior of a surface revolved around the relationship between the extension of interior for a depicted surface, and the extension of inclusion for regions which depict the surfaces. These intuitions are formalized in the definition of the class of structures \( \mathcal{M}_i^{\text{region}} \). Intuition 16 is formalized by conditions (4) and (5) in the definition of \( \mathcal{M}_i^{\text{region}} \); these conditions constrain the relationship between the patches in the edge-position graph of a surface (which determine the extension of interior) and the planar subdivisions in the depicting planar subgraph (which determine the extension of inclusion). In each of these conditions, we are constraining the extension of the depiction relation \( \Delta \) for surfaces and regions. Condition (6) formalizes Intuition 17; it completely specifies the conditions under which the extensions of interior and inclusion will be isomorphic for a region depicting the
surface. Similarly, condition (7) formalizes Intuition 18, by completely specifying the conditions under which the extensions of hole\_interior and inclusion will be disjoint for a region depicting the surface.

**Definition 6.15** Let $\mathcal{M}_i^\text{region}$ be the following class of structures:

1. $\mathcal{M}$ is an extension of a structure in $\mathcal{M}_i^\text{interior}$ and a structure in $\mathcal{M}_i^\Delta$.
2. $\mathcal{M}$ is an extension of a structure in $\mathcal{M}_i^\text{planar}$.
3. The set of points in $\mathcal{M}$ is position-isomorphic to the set of pixels in $\mathcal{M}$. (Intuitively, position is preserved by depiction.)
4. Within a depicted interior graph for a surface $s$, there exists a position which is an element of a patch in the edge-position graph of $s$ and which is also an element of a planar subdivision in the depicting planar subgraph.
5. If a position is both an element of a patch in the edge-position graph of a surface and it is an element of a planar subdivision in the planar image graph, then there exists a depicted interior graph in $\mathcal{M}$ which contains both the patch and the planar subdivision.
6. For each surface $s \in \mathcal{M}$, the planar subdivisions in the depicted interior graph are isomorphic to the patches in the edge-position graph of the surface if the surface substructure for $s$ is isomorphic to the region substructure for the region in the depicted interior subgraph.
7. For each surface $s \in \mathcal{M}$, the planar subdivisions in the depicted hole graph are disjoint from the patches in the hole-position graphs of the surface if the surface substructure for $s$ which contains only hole edges is a subgraph of the region substructure for the region in the depicted hole subgraph.

We can illustrate the structures in $\mathcal{M}_i^\text{region}$ using the images in Figure 6.13.

**Example:** Let $I$ be the image in Figure 6.13(a). In this image, there is a triangular surface with a hole.

Let $F_1$ be the face in the planar subdivision corresponding to the region $r_1$, and let $F_2$ be the face in the planar subdivision corresponding to the region $r_2$.

There exists a structure $\mathcal{M} \in \mathcal{M}_i^\text{region}$ such that $\mathcal{M} \models I$ and which has the following properties:

$\langle r_1, s_1 \rangle \in \Delta$

$\langle l_1, e_1 \rangle, \langle l_2, e_2 \rangle, \langle l_3, e_3 \rangle, \langle l_4, e_4 \rangle, \langle l_5, e_5 \rangle, \langle l_6, e_6 \rangle \in \Delta$

$\langle e_1, s_1 \rangle, \langle e_2, s_1 \rangle, \langle e_3, s_1 \rangle, \langle e_4, s_1 \rangle, \langle e_5, s_1 \rangle, \langle e_6, s_1 \rangle \in \text{part}$

The surface substructure is shown in Figure 6.14(b), and the region substructure is shown in Figure 6.14(a). The depicted interior graph is shown in Figure 6.15: the arc between $r_1$ and $F_1$ indicates that the region $r_1$ corresponds to the face $F_1$ in the planar subdivision. The arc between $\Pi_1$ and $F_1$ indicates that the set of positions in the patch $\Pi_1$ is a subset of the set of positions in the face $F_1$.

In this structure, the set of positions in the face $F_1$ is equal to the union of the positions in the patches of the edge-position graph, so that $F_1 = \Pi_1 \cup \Pi_2 \cup \Pi_3$. Since the region substructure for
Figure 6.13: Images used for examples of structures in $\mathcal{M}_r^{\text{region}}$. 
Figure 6.14: Surface and region substructures in the structure corresponding to the image in Figure 6.13(a).

Figure 6.15: Depicted interior graph in the structure corresponding to the image in Figure 6.13(a).
$r_1$ is isomorphic to the surface substructure for the surface $s_1$, condition (6) in the definition of $M_i^{\text{region}}$ is satisfied by $\mathcal{M}$.

The depicted hole graph is shown in Figure 6.16; the arc between $r_2$ and $F_2$ indicates that the region $r_2$ corresponds to the face $F_2$ in the planar subdivision. The arc between $\Pi_4$ and $F_2$ indicates that the set of positions in the patch $\Pi_4$ is a subset of the set of positions in the face $F_2$.

In this structure, the set of positions in the face $F_2$ is equal to the set of positions in the patch of the hole-position graph, and hence it is disjoint from the set of positions in the face corresponding to the region $r_1$. Since the surface substructure for the surface $s_1$ consisting only of hole edges $(e_4, e_5, e_6)$ is isomorphic to a subgraph the region substructure for $r_2$, condition (7) in the definition of $M_i^{\text{region}}$ is satisfied by $\mathcal{M}$.

Example: Let $I$ be the image in Figure 6.13(c). Intuitively, in this image we have two overlapping
Figure 6.18: Depicted interior graph in the structure corresponding to the image in Figure 6.13(c).

surfaces, but the occluding edges are not depicted because of errors in edge detection.

Let $F_1$ be the face in the planar subdivision corresponding to the region $r_1$. There exists a structure $\mathcal{M} \in \mathcal{M}_{\text{region}}^r$ such that $\mathcal{M} \models I$ and which has the following properties:

$\langle r_1, s_1 \rangle, \langle r_1, s_2 \rangle \in \Delta$

$\langle l_1, e_1 \rangle, \langle l_2, e_6 \rangle, \langle l_3, e_5 \rangle, \langle l_4, e_8 \rangle, \langle l_5, e_7 \rangle, \langle l_6, e_2 \rangle, \langle l_7, e_3 \rangle, \langle l_8, e_4 \rangle \in \Delta$

$\langle e_1, s_1 \rangle, \langle e_2, s_1 \rangle, \langle e_3, s_1 \rangle, \langle e_4, s_1 \rangle, \langle e_5, s_2 \rangle, \langle e_6, s_2 \rangle, \langle e_7, s_2 \rangle, \langle e_8, s_2 \rangle \in \text{part}$

The surface substructure is shown in Figure 6.17(a), and the region substructure is shown in Figure 6.17(b). The depicted interior graph is shown in Figure 6.18: the arc between $r_1$ and $F_1$ indicates that the region $r_1$ corresponds to the face $F_1$ in the planar subdivision. The arc between $\Pi_1$ and $F_1$ indicates that the set of positions in the patch $\Pi_1$ is a subset of the set of positions in the face $F_1$.

In this structure, the surface substructure for either $s_1$ or $s_2$ is not isomorphic to the region substructure for $r_1$. Consequently, the planar subdivision $F_1$ is not isomorphic to the edge-position graph of either $s_1$ or $s_2$ alone, since $F_1 = \Pi_1 \cup \Pi_2$.

However, since $\Pi_1 \subset F_1$, there do exist positions which are elements of a patch in the edge-position graph of $s_1$ and which is also an element of the planar subdivision $F_1$ in the depicting planar subgraph. Thus, $\mathcal{M}$ in this example satisfies condition (4) in the definition of $\mathcal{M}_{\text{region}}^r$. $\square$
Figure 6.19: Region substructure in the structure corresponding to the image in Figure 6.13(b).

Figure 6.20: Depicted interior graph and depicted hole graph in the structure corresponding to the image in Figure 6.13(b).
Example: Let $I$ be the image in Figure 6.13(b). In this image, there is a triangular surface with a hole, but one of the hole edges is not depicted due to errors in edge detection.

Let $F_1$ be the face in the planar subdivision corresponding to the region $r_1$.

There exists a structure $\mathcal{M} \in \mathcal{M}_i^{region}$ such that $\mathcal{M} \models I$ and which has the following properties:

$$\langle r_1, s_1 \rangle \in \Delta$$

$$\langle l_1, e_1 \rangle, \langle l_2, e_2 \rangle, \langle l_3, e_3 \rangle, \langle l_4, e_4 \rangle, \langle l_5, e_5 \rangle \in \Delta$$

$$\langle e_1, s_1 \rangle, \langle e_2, s_1 \rangle, \langle e_3, s_1 \rangle, \langle e_4, s_1 \rangle, \langle e_5, s_1 \rangle, \langle e_6, s_1 \rangle \in \text{part}$$

The surface substructure is the same as the one in Figure 6.14(b), since we are dealing with the same surface. The region substructure is shown in Figure 6.19. The depicted interior graph and the depicted hole graph are shown in Figure 6.20. The arc between $r_1$ and $F_1$ indicates that the region $r_1$ corresponds to the face $F_1$ in the planar subdivision. The arcs between the patches $\Pi_i$ and $F_1$ indicate that the sets of positions in each patch $\Pi_i$ is a subset of the set of positions in the face $F_1$: together we have $F_1 = \Pi_1 \cup \Pi_2 \cup \Pi_3 \cup \Pi_4$.

Since the set of positions in $F_1$ is not equal to the patches in the edge-position graph, if condition (6) in the definition of $\mathcal{M}_i^{region}$ is to be satisfied by $\mathcal{M}$, then the region substructure for $r_1$ is not isomorphic to the surface substructure for the surface $s_1$, which is the case.

Similarly, the set of positions in the face $F_1$ is not equal to the set of positions in the patch of the hole-position graph, and hence it is not disjoint from the set of positions in the face corresponding to the region $r_1$. If condition (7) in the definition of $\mathcal{M}_i^{region}$ is to be satisfied by $\mathcal{M}$, the surface substructure for the surface $s_1$ consisting only of hole edges $(e_4, e_5, e_6)$ cannot be isomorphic to a subgraph the region substructure for $r_2$, which is the case. □

**Theorem 6.5 Existence Theorem for $\mathcal{M}_i^{region}$**

*The class of structures $\mathcal{M}_i^{region}$ exists and is nonempty.*

**Proof:** Since structures in $\mathcal{M}_i^{region}$ are the union of the following substructures, we will first need to show that these classes of graphs exist:

1. edge-position graphs
2. hole-position graphs
3. position-isomorphic sets
4. depicting planar subgraphs
5. depicted interior graphs
6. depicted hole graphs

By the Existence Theorem for $\mathcal{M}_i^{interior}$, we know that edge-position and hole-position graphs exist.

Position-isomorphic sets exist, since they are subsets of the the points and pixels within the structure.
Depicting planar subgraphs are arbitrary subgraphs of planar graphs, and hence they exist. Since, depicted interior graphs are the union of arbitrary edge-position graphs and depicting planar subgraphs, they exist. Similarly, depicted hole graphs are the union of arbitrary hole-position graphs and depicting planar subgraphs.

The remaining conditions in the definition of $\mathcal{M}_{\text{interior}}$ are simply restricting which substructures are being combined. Thus, conditions (4) and (5) specify that the edge-position and depicting planar subgraphs must have positions in common. The existence of such positions is guaranteed by the position-isomorphic sets.

Conditions (6) and (7) allow combination of arbitrary subgraphs; they merely specify the conditions under which the subgraphs must be isomorphic.

### 6.3.3 Axiomatization of Depiction for \textit{interior}

We will denote by $\Delta_{\text{interior}}$ the module which axiomatizes the structures in $\mathcal{M}_{\text{region}}$ (see Figure 6.21).

There is a close correspondence between the axioms of $\Delta_{\text{interior}}$ and the intuitions we earlier presented. Intuition 15 corresponds to axiom 6.10. Intuition 16 leads to axioms 6.11 and axiom 6.12. Axiom 6.11 requires that for any depicted surface there must exist some position which is both interior to the surface and included in the depicting region. Axiom 6.12 can be considered to be a partial converse - if a position is both interior to a surface and included in a region, then there exists some surface depicted by the region such that the position is also interior to this surface. We cannot have a full converse to this axiom because of possible occlusion in the scene. For example, there may be a occluded surface underneath the surface depicted by the region, so that we cannot simply conclude that the region depicts a surface simply because there exists a position interior to the occluded surface.

Earlier intuitions indicated that the depiction relation between \textit{interior} and \textit{inclusion} is not one of equivalence: in particular, there can possibly be positions interior to some surface that are not included in the region that depicts the surface. The depiction axiom for \textit{interior} (axiom 6.13 of $\Delta_{\text{interior}}$) captures the conditions under which the extensions of these two relations are equivalent. This axiom has a disjunctive consequent which specifies the conditions under which positions interior to a surface are equivalent to positions included in the depicting region. In the first conjunct in the consequent of axiom 6.13, a position is included in a region and is interior to a surface depicted by that region if every border line of the region depicts a boundary edge of the surface. This can be illustrated by the images in Figure 6.11(b) and (d).

**Example:** Let $I$ be the image Figure 6.11(d) where

$$I \models \text{inclusion}(r_2, t_1) \land \text{border}(l_1, r_1) \land \text{border}(l_2, r_1)$$

There is a model $\mathcal{M}$ of $\Delta_{\text{interior}} \cup T_{\text{interior}} \cup T_{fg} \cup \Delta \cup T_{\text{kernel}} \cup T_{\text{scene}} \cup I$ such that

$$\langle r_1, s_1 \rangle, \langle r_2, s_2 \rangle \in \Delta$$

$$\langle l_1, e_1 \rangle, \langle l_2, e_2 \rangle \in \Delta$$
Since
\[ \langle e_1, s_2 \rangle, \langle e_2, s_2 \rangle \in \text{part} \]
there exist border lines of \( r_1 \) which do not depict boundary edges in \( s_1 \). Since
\[ \langle s_1, t_1 \rangle \in \text{interior} \]
axiom 6.13 gives
\[ \langle r_1, t_1 \rangle \notin \text{inclusion} \]
\( \Box \)

In the second conjunct in the consequent of axiom 6.13, a position is interior to a surface and is included in a region depicting the surface if every depicted point in the surface is contained in a border line of the region. This can be illustrated by the image in Figure 6.11(a).

**Example:** Let \( I \) be the image in Figure 6.11(a) where
\[ I \models \text{inclusion}(r_1, t_1) \land \text{inclusion}(r_1, t_2) \land \text{inclusion}(r_1, t_3) \]

There is a model \( \mathcal{M} \) of \( T^\Delta_{\text{interior}} \cup T_{\text{interior}} \cup T_f \cup T_\Delta \cup T_{\text{kernel}} \cup T_{\text{scene}} \cup I \) such that
\[ \langle r_1, s_1 \rangle, \langle r_1, s_2 \rangle \in \Delta \]
\[ \langle h_1, e_1 \rangle \in \Delta, \langle e_1, s_2 \rangle \in \text{part} \]
\[ \langle s_1, t_1 \rangle, \langle s_2, t_2 \rangle \in \text{interior} \]
\[ \langle p_1, s_1 \rangle \in \text{part}, \langle p_1, t_3 \rangle \in \text{position} \]

Since the point \( p_1 \) is not depicted by a pixel in a border line for \( r_1 \), axiom 6.13 gives
\[ \langle s_1, t_2 \rangle, \langle s_2, t_1 \rangle \notin \text{interior} \]

Also note that in this structure, there are border lines of \( r_1 \) which do not depict boundary edges of \( s_1 \). \( \Box \)

We can also think of axiom 6.13 in terms of its contrapositive – if there is a border line which does not depict a boundary edge of the surface, then there will exist a position which is interior to the surface yet not included in the depicting region. Also, if there is a depicted point in the surface which is not contained in a border line of the region, then there will exist a position which is included in the region but which is not interior to the surface.

Similarly, earlier intuitions indicated that the extension of \( \text{hole.interior} \) and \( \text{inclusion} \) is not always disjoint: in particular, there can possibly be positions interior to a hole in some surface that are included in the region that depicts the surface. The depiction axiom for \( \text{hole.interior} \) (axiom 6.14 of \( T^\Delta_{\text{interior}} \))
Depiction preserves the position of depicted points.

\[(\forall p, q, t) \text{position}(p) \land \Delta(q, p) \supset ((\text{position}(p, t) \equiv \text{position}(q, t))\]  \hspace{1cm} (6.10)

If a region depicts a surface, then there exists a position which is both interior to the surface and included in the region.

\[(\forall r, s) \Delta(r, s) \supset (\exists t) \text{interior}(s, t) \land \text{inclusion}(r, t)\]  \hspace{1cm} (6.11)

If a position is both interior to a surface and included in a region, then there exists a surface depicted by the region such that the position is also interior to the depicted surface.

\[(\forall r, s_1, t) \text{interior}(s_1, t) \land \text{inclusion}(r, t) \supset (\exists s_2) \Delta(r, s_2) \land \text{interior}(s_2, t)\]  \hspace{1cm} (6.12)

For any depicted surface, the extension of interior is equivalent to the extension of inclusion iff every border line in the region depicts a boundary edge of the surface and every depicted point is contained in a border line of the region.

\[(\forall r, s, t) \Delta(r, s) \supset ((\text{inclusion}(r, t) \equiv \text{interior}(s, t)) \equiv \\
(\forall l) \text{border}(l, r) \supset (\exists e) \Delta(l, e) \land \text{boundary}(e, t) \land \text{part}(e, s)) \land ((\forall p, q) \text{part}(p, s) \land \Delta(q, p) \supset (\exists l) \text{border}(l, r) \land \text{in}(q, l))\]  \hspace{1cm} (6.13)

For any depicted surface, the extension of hole_interior is disjoint from the extension of inclusion iff every depicted point in a hole boundary edge is contained in a border line of the region.

\[(\forall r, s, t) \Delta(r, s) \supset ((\text{inclusion}(r, t) \equiv \neg \text{hole_interior}(s, t)) \equiv \\
(\forall e, q, p) \text{part}(p, e) \land \text{part}(e, s) \land \text{hole_boundary}(e, t) \land \Delta(q, p) \supset (\exists l) \text{border}(l, r) \land \text{in}(q, l))\]  \hspace{1cm} (6.14)

\text{Figure 6.21: } T_{\text{interior}}^\Delta: \text{Depiction axioms for the interior of a surface.}
captures the conditions under which the extensions of these two relations are disjoint - they are disjoint if and only if every depicted point in a hole boundary edge is contained in a border line of the region depicting the surface. We can again consider the contrapositive of this axiom - if there exists a depicted point in a hole boundary edge which is not contained in a border line of the depicting region, then there exists a position which is interior to the hole and also included in the region. Intuitively, this means that there is a piece "missing" in the line depicting the hole boundary edge. These intuitions are illustrated by the images in Figure 6.12.

Example: Let $I$ be the image in Figure 6.12(a), where

$$I \models \text{border}(l_1, r_1) \land \text{border}(l_2, r_1) \land \text{border}(l_3, r_1) \land \text{border}(l_4, r_1) \land \text{border}(l_5, r_1) \land \text{border}(l_6, r_1) \land \text{border}(l_7, r_1) \land \text{border}(l_8, r_1) \land \text{border}(l_9, r_2) \land \text{border}(l_{10}, r_2)$$

There is a model $M$ of $T^A_{\text{interior}} \cup T_{\text{interior}} \cup T_{\text{surface}} \cup T_{\text{kernel}} \cup T_{\text{scene}} \cup I$ such that $\{e_5, e_6, e_7, e_8, s\} \subseteq M$ and

$$\langle r_1, s_1 \rangle \in \Delta$$

$$\langle l_5, e_5 \rangle, \langle l_6, e_6 \rangle, \langle l_7, e_7 \rangle, \langle l_8, e_8 \rangle \in \Delta$$

$$\langle s_1, t_1 \rangle \in \text{hole_interior}$$

$$\langle r_1, t_1 \rangle \notin \text{inclusion}$$

In this model, the region $r_2$ does not depict a surface, but rather is the hole in the surface $s$. Since the hole boundary edges for the surface $s$ are all depicted, and each point in these edges is depicted by a pixel in a border line of both $r_1$ and $r_2$, by axiom 6.14, the extensions of $\text{inclusion}$ and $\text{hole_interior}$ are disjoint. □

Example: Let $I$ be the image in Figure 6.12(b), where

$$I \models \text{border}(l_1, r_1) \land \text{border}(l_2, r_1) \land \text{border}(l_3, r_1) \land \text{border}(l_4, r_1)$$

There is a model $M$ of $T^A_{\text{interior}} \cup T_{\text{interior}} \cup T_{\text{surface}} \cup T_{\text{kernel}} \cup T_{\text{scene}} \cup I$ such that $\{e_5, e_6, e_7, e_8, p_1, s\} \subseteq M$ and

$$\langle r_1, s_1 \rangle \in \Delta$$

$$\langle p_1, e_7 \rangle \in \text{part}, \langle p_1, t_1 \rangle \in \text{position}$$

$$\langle l_5, e_5 \rangle, \langle l_6, e_6 \rangle, \langle l_7, e_7 \rangle, \langle l_8, e_8 \rangle \in \Delta$$

$$\langle s_1, t_1 \rangle \in \text{hole_interior}$$

$$\langle r_1, t_1 \rangle, \langle r_1, t_2 \rangle \in \text{inclusion}$$

Since there exists a point in a hole boundary edges that is depicted by a pixel but which is not in a border line of $r_1$, by axiom 6.14, the extensions of $\text{inclusion}$ and $\text{hole_interior}$ are not disjoint. □
6.3.4 Characterization Theorems for $T_{\text{interior}}^\Delta$

In this section, we show that the structures in $\mathcal{M}^\text{region}_i$ satisfy the axioms of $T_{\text{interior}}^\Delta$, and that the models of $T_{\text{interior}}^\Delta$ correspond to the structures in $\mathcal{M}^\text{region}_i$.

Satisfiability of $T_{\text{interior}}^\Delta$

Theorem 6.6 Every structure in $\mathcal{M}^\text{region}_i$ is a model of $T_{\text{kernel}} \cup T_{fg} \cup T_{\Delta} \cup T_{\text{interior}} \cup T_{\text{interior}}^\Delta \cup I$ for some planar image $I$.

Proof: It is easy to see that axiom 6.10 of $T_{\text{interior}}^\Delta$ is satisfied by $\mathcal{M}$ from property (3) in the definition of $\mathcal{M}^\text{region}_i$.

Consider axiom 6.11 of $T_{\text{interior}}^\Delta$:

\[ \mathcal{M} \models (\forall r, s) \Delta(r, s) \supset (\exists t) \text{interior}(s, t) \land \text{inclusion}(r, t) \]

iff for any variable assignment $\sigma$, and every region $\sigma(r)$ and surface $\sigma(s)$, we have

\[ \mathcal{M}, \sigma \models \Delta(r, s) \supset (\exists t) \text{interior}(s, t) \land \text{inclusion}(r, t) \]

This is equivalent to saying that if

\[ \langle \sigma(r), \sigma(s) \rangle \in \Delta \]

then there exists a position $t \in M$ such that

\[ \langle \sigma(s), t \rangle \in \text{interior} \text{ and } \langle \sigma(r), t \rangle \in \text{inclusion}. \]

By hypothesis, there exists a depicted interior graph for the surface $\sigma(s)$. By property (4) in the definition of $\mathcal{M}^\text{region}_i$, there exists a position which is an element of a patch in the edge-position graph of $s$, so that

\[ \langle \sigma(s), t \rangle \in \text{interior} \]

By property (4) this position $t$ is also an element of a planar subdivision in the depicting planar subgraph, so that

\[ \langle \sigma(r), t \rangle \in \text{inclusion} \]

by which the axiom is satisfiable.

Consider axiom 6.12 of $T_{\text{interior}}^\Delta$:

\[ \mathcal{M} \models (\forall r, s_1, t) \text{interior}(s_1, t) \land \text{inclusion}(r, t) \supset (\exists s_2) \Delta(r, s_2) \land \text{interior}(s_2, t) \]

iff for any variable assignment $\sigma$, and for every region $\sigma(r) \in M$, every surface $\sigma(s_1) \in M$, and every position $\sigma(t) \in M$, if

\[ \mathcal{M}, \sigma \models \text{interior}(s_1, t) \land \text{interior}(s, t) \]

then there exists a surface $s_2 \in M$ such that

\[ \langle \sigma(r), s_2 \rangle \in \Delta, \langle s_2, \sigma(t) \rangle \in \text{interior} \]

The antecedent of this sentence holds iff the position $\sigma(t)$ is both an element of a patch in the edge-position graph of a surface and it is an element of a planar subdivision in the planar image graph.

The consequent of this sentence holds iff there exists a depicted interior graph in $\mathcal{M}$ which contains both the path and the planar subdivision. Thus the axiom follows from condition (3) in the definition of $\mathcal{M}^\text{region}_i$. 

Consider axiom 6.13 of $T^\Delta_{\text{interior}}$: Suppose
\[ M, \sigma \models (\forall r, s, t) \Delta(r, s) \supset [(\text{inclusion}(r, t) \equiv \text{interior}(s, t)) \equiv \\
(\forall l) \text{border}(l, r) \supset (\exists e) \Delta(l, e) \land \text{boundary}(e, t) \land \text{part}(e, s) \land \\
\land ((\forall p, q) \text{part}(p, s) \land \Delta(q, p) \supset (\exists l) \text{border}(l, r) \land \text{in}(q, l))] \]

This is equivalent to saying that for any variable assignment $\sigma$, the extension of $\text{inclusion}$ is equivalent to the extension of $\text{interior}$ for a surface $\sigma(s)$ depicted by the region $\sigma(r)$ iff
\[ M, \sigma \models (\forall l) \text{border}(l, r) \supset (\exists e) \Delta(l, e) \land \text{boundary}(e, t) \]
and
\[ M, \sigma \models ((\forall p, q) \text{part}(p, s) \land \Delta(q, p) \supset (\exists l) \text{border}(l, r) \land \text{in}(q, l))] \]

Since by hypothesis, the extension of $\text{inclusion}$ for $\sigma(r)$ is isomorphic to the extension of $\text{interior}$ for $\sigma(s)$, by condition (6) in the definition of $M^\text{region}$, the surface substructure for $\sigma(s)$ must be isomorphic to the border substructure for $\sigma(r)$. By the isomorphism, every line in the border substructure must depict an edge in the surface substructure, so that the first formula above is satisfied.

Similarly, by the isomorphism, every point in the surface substructure is depicted by a pixel in the border substructure, so the second formula above is satisfied.

Thus, axiom 6.13 is satisfied by $M$.

Finally, consider axiom 6.14 of $T^\Delta_{\text{interior}}$. Suppose
\[ M \models \\
(\forall r, s, t) \Delta(r, s) \supset [(\text{inclusion}(r, t) \equiv \neg \text{hole_interior}(s, t)) \equiv \\
(\forall e, q, p) \text{part}(p, e) \land \text{part}(e, s) \land \neg \text{hole_interior}(e, t) \land \Delta(q, p) \supset (\exists l) \text{border}(l, r) \land \text{in}(q, l))] \]

This is equivalent to saying that for any variable assignment $\sigma$, the extension of $\text{inclusion}$ is disjoint from the extension of $\text{hole_interior}$ for a surface $\sigma$ depicted by the region $\sigma(r)$ iff
\[ M, \sigma \models ((\forall e, q, p) \text{part}(p, e) \land \text{part}(e, s) \land \neg \text{hole_interior}(e, t) \land \Delta(q, p) \supset (\exists l) \text{border}(l, r) \land \text{in}(q, l))] \]

Since by hypothesis, the extension of $\text{inclusion}$ for $\sigma(r)$ is disjoint from the extension of $\text{hole_interior}$ for $\sigma(s)$, by condition (7) in the definition of $M^\text{region}$, every depicted point in a hole boundary edge of the surface is a pixel in the boundary of the region in the depicted interior subgraph. It is easy to see that the consequent of the axiom is satisfied in $M$. \qed

**Axiomatizability of $T^\Delta_{\text{interior}}$**

**Theorem 6.7** Let $M$ be a structure in $L_{\text{interior}} \cup L_\Delta \cup L_{\text{planar}}$ with the following properties:

- $M$ contains a substructure isomorphic to a structure in $M^\text{polygon}_i$;
- $M$ contains a substructure isomorphic to a structure in $M^\text{single}_i$;
- $M$ contains a substructure isomorphic to a structure in $M^\text{plane}_i$;
- $M$ contains a substructure isomorphic to a structure in $M^\text{planar}_i$;
- $M$ is a model of $T_{\text{kernel}} \cup T_f \cup T_\Delta \cup T_{\text{interior}} \cup T_{\text{interior}}^{\Delta} \cup I$ for some planar image $I$. 
Then \( \mathcal{M} \) is isomorphic to a structure in \( \mathcal{M}_i^{\text{region}} \).

**Proof:** Let \( \mathcal{M} \) be a model of \( T_{\text{kernel}} \cup T_{f_g} \cup T_{\Delta} \cup T_{\text{interior}} \cup T_{\text{interior}}^{\Delta} \) which satisfies the properties in the hypothesis of the theorem.

It is easy to see that by axiom 6.10, the set of points in \( \mathcal{M} \) is position-isomorphic to the set of pixels in \( \mathcal{M} \), so that condition (3) in the definition of \( \mathcal{M}_i^{\text{region}} \) is satisfied by \( \mathcal{M} \).

Since \( \mathcal{M} \) is a model of \( T_{\text{scene}} \cup T_{f_g} \cup T_{\text{interior}} \), it is an extension of a structure in \( \mathcal{M}_i^{\text{interior}} \), and hence there is an edge-position graph and a hole-position graph within \( \mathcal{M} \) for each surface \( s \in \mathcal{M} \). Similarly, by hypothesis, \( \mathcal{M} \) is an extension of a structure in \( \mathcal{M}_i^{\text{planar}} \), and hence there is a planar subdivision within \( \mathcal{M} \) for each region \( r \in \mathcal{M} \).

By axiom 6.11, for every region \( r \) which depicts a surface \( s \), there exists a position which is an element of a patch in the edge-position graph of \( s \) and which is also an element of a planar subdivision. Thus, condition (4) in the definition of \( \mathcal{M}_i^{\text{region}} \) is satisfied by \( \mathcal{M} \).

By axiom 6.12,

\[
\mathcal{M} \models (\forall r, s, t) \text{ interior}(s, t) \land \text{inclusion}(r, t) \supset (\exists s_2) \Delta(r, s_2) \land \text{interior}(s_2, t)
\]

iff for any variable assignment \( \sigma \), and any elements \( \sigma(r), \sigma(s_1), \sigma(t) \in \mathcal{M} \), if

\[
\mathcal{M}, \sigma \models \text{interior}(s_1, t) \land \text{inclusion}(r, t)
\]

then there exists a surface \( s_2 \in \mathcal{M} \) such that

\[
\langle \sigma(r), s_2 \rangle \in \Delta, \langle s_2, \sigma(t) \rangle \in \text{interior}
\]

The antecedent is true iff the position \( \sigma(t) \) is an element of a patch in the edge-position graph for \( \sigma(s_1) \) and it is also an element of a planar subdivision in the planar image graph for \( \sigma(r) \). The antecedent is then true if the position \( \sigma(t) \) is an element of a patch in the depicted interior graph for \( s_2 \). Thus, condition (5) in the definition of \( \mathcal{M}_i^{\text{region}} \) is satisfied by \( \mathcal{M} \).

By axiom 6.13, if

\[
\mathcal{M} \models (\forall r, s, t) \Delta(r, s) \supset [(\text{inclusion}(r, t) \equiv \text{interior}(s, t))]
\]

then

\[
\mathcal{M} \models ((\forall l) \text{ border}(l, r) \supset (\exists e) \Delta(l, e) \land \text{boundary}(e, t) \land \text{part}(e, s))
\]

\[
\land ((\forall p, q) \text{ part}(p, s) \land \Delta(q, p) \supset (\exists l) \text{ border}(l, r) \land \text{in}(q, l))
\]

By the antecedent, for any variable assignment \( \sigma \), the extension of \( \text{interior} \) for the surface \( \sigma(s) \) is isomorphic to the extension of \( \text{inclusion} \) for the region \( \sigma(r) \); that is, the patches in the edge-position graph for \( \sigma(s) \) is isomorphic to the planar subdivisions in the depicted interior graph.

Suppose that the surface substructure for \( \sigma(s) \) is not isomorphic to the border substructure for \( \sigma(r) \). Then one of the following conditions must hold:

- There exists line in the border substructure which does not depict an edge in the surface substructure.
- There exists an edge (point) in the surface substructure which is not depicted by a line (pixel) in the border substructure.

However, the first condition is false by the first conjunct in the antecedent of axiom 6.13, and the second condition is false by the second conjunct of axiom 6.13. Therefore, the surface substructure for \( \sigma(s) \) must be isomorphic to the border substructure for \( \sigma(r) \), and hence condition (6) in the definition of \( \mathcal{M}_i^{\text{region}} \) must be satisfied by \( \mathcal{M} \).

By axiom 6.14, if

\[
\mathcal{M} \models (\forall r, s, t) \Delta(r, s) \supset [(\text{inclusion}(r, t) \equiv \neg \text{hole.interior}(s, t))]
\]
then
\[ M \models \land ((\forall e. q. p) \text{part}(p, e) \land \text{part}(e, s) \land \text{hole-boundary}(e, t) \land \Delta(q, p) \supset (\exists l) \text{border}(l, r) \land \text{in}(q, l))] \]

By the antecedent, for any variable assignment \(\sigma\), the extension of \(\text{hole-interior}\) for the surface \(\sigma(s)\) is disjoint from the extension of \(\text{inclusion}\) for the region \(\sigma(r)\); that is, the set of patches in the hole-position graph for \(\sigma(s)\) is disjoint from the planar subdivisions in the depicted interior graph. By the antecedent of axiom 6.14, every depicted point in a hole boundary edge of the surface is a pixel in the boundary of the region, and hence condition (7) in the definition of \(M_{\text{region}}\) must be satisfied by \(M\). \(\Box\)
Chapter 7

A Theory of Occlusion

To this point, we have introduced six sets of axioms, which have been used to prove several characterization theorems about shape and depiction. In this chapter we introduce the two final sets of axioms in the CardWorld theories — $T_{obsures}$ and $T_{occ}$ — which together will constitute a theory of occlusion. Intuitively, a scene object is occluded when it is not depicted due to a surface that is on top of it. With the occlusion axioms, we want to make this intuition precise. The module $T_{obsures}$ will formalize intuitions about surfaces being on top of one another; it will be a theory containing only scene relations. The module $T_{occ}$ will formalize the relationship between these scene relations and nondepiction of scene objects within an image.

7.1 Obscuring Surfaces

In this section, we axiomatize the notion of how surfaces can possibly obscure each other within a scene. In doing so, we introduce the CardWorld module $T_{obsures}$ and consider the theory $T_{obsures} \cup T_{interior} \cup T_{fg} \cup T_{scene}$. It is important to realize that this theory considers scene relations alone. In the next section, we will combine the concepts of obscuring surfaces with depiction in order to provide a complete characterization of occlusion.

Our first task will be to specify the conditions that make occlusion possible — what is it about a surface that allows it to occlude other scene objects? A scene object is occluded when there is a surface on top of the object; this leads to the requirement that a surface occludes another scene object because it constitutes a figure. In this way, the axioms for the interior of surfaces from the previous chapter will play a crucial role.

Closely related to this is the intuition that when one surface occludes another, there is a boundary, or occluding edge, at the point of occlusion that defines the figure of the occluding surface. Part of the characterization theorem for $T_{obsures}$ will include conditions on the existence of occluding edges. Similar intuitions can be found in [Williams and Hanson 96].
7.1.1 Intuitions for Obscuring Surfaces

The nonlogical lexicon for $T_{\text{obsures}}$ is

$$L_{\text{obsures}} = L_{\text{interior}} \cup \{\text{obsures}(s_1, s_2, t), \text{abuts}(e_1, e_2, t), \text{ontop}(s_1, s_2), \text{partial-occludes}(s_1, s_2), \text{total-occludes}(s_1, s_2), \text{occluding}(e, s)\}$$

- The ternary predicate $\text{obsures}(s_1, s_2, t)$ denotes the relation which represents one surface $s_1$ which obscures, or is on top of, surface $s_2$ at the position $t$.

- The ternary predicate $\text{abuts}(e_1, e_2, t)$ is a relation over edges in different surfaces and a position. Intuitively, it represents two surfaces which are abutting each other at some position in a scene along two of their edges.

- The binary predicate $\text{ontop}(s_1, s_2)$ over surfaces denotes that surface $s_1$ is on top of surface $s_2$ within the scene.

- The binary predicate $\text{partial-occludes}(s_1, s_2)$ over surfaces denotes that surface $s_1$ obscures only part of $s_2$.

- The binary predicate $\text{total-occludes}(s_1, s_2)$ denotes that surface $s_1$ totally obscures $s_2$, that is, $s_1$ obscures $s_2$ at every position interior to $s_2$.

- The binary predicate $\text{occluding}(e, s)$ denotes that the edge $e$ is occluding the surface $s$ at some position.

The following examples illustrate some of the possible scenes which correspond to our intuitions about obscuring surfaces. It is important to note that $\text{obsures}$ is a scene relation that is independent of depiction. In all of the examples given in this chapter, we will be referring to properties of scenes. In any of the diagrams containing scenes, if a scene object is not displayed in the figure, then it should be interpreted as being occluded. We may sometimes want to explicitly refer to the occluded scene objects: in these cases, we will use diagrams that display the hidden or occluded edges, and they will be explicitly referred to in this way. It should be stressed that none of these diagrams are images: they are pictorial representations of scenes.

**Intuition 19** Surfaces can only be occluded at positions interior to the occluding surface.

**Intuition 20** The surfaces in a scene are ordered according to their "depth" at different positions in the scene. Further, in any scene, there is always a surface which is on top of every other surface at some position.

**Intuition 21** Surfaces can be partially occluded or totally occluded by other surfaces in the scene.
Figure 7.1: Scene used to illustrate intuitions for obscuring surfaces.
Example: Consider the scene in Figure 7.1(a), in which dashed lines represent displayed hidden edges of the surfaces. There exists a structure $\mathcal{M} \in \mathcal{L}_{\text{obscures}}$ such that
\[ \{s_1, s_2, s_3, s_4, t_1, t_2, t_3, t_4, t_5\} \subseteq \mathcal{M} \text{ and which has the following properties:} \]

Intuitively, we have
\[ \langle s_1, s_2, t_1 \rangle, \langle s_1, s_2, t_2 \rangle, \langle s_2, s_3, t_2 \rangle, \langle s_1, s_3, t_2 \rangle \in \text{obscures} \]

However, we do have
\[ \langle s_2, s_3, t_1 \rangle, \langle s_3, s_2, t_1 \rangle \notin \text{obscures} \]

since the position $t_1$ is not interior to the surface $s_3$.

In this case, the surfaces $s_2$ and $s_3$ are both partially occluded by the surface $s_1$. In fact, we can order the surfaces according to their "depth" in the scene — $s_1$ is on top of all other surfaces, and $s_2$ is on top of $s_3$.

Also note that we have
\[ \langle s_3, s_4, t_5 \rangle \in \text{abuts} \]

so that neither $s_3$ nor $s_4$ are on top of each other. \(\square\)

Example: Consider the scene in Figure 7.1(b), in which dashed lines represent displayed hidden edges of the surfaces. There exists a structure $\mathcal{M} \in \mathcal{L}_{\text{obscures}}$ such that $\{s_1, s_2\} \subseteq \mathcal{M}$ in which the surface $s_2$ is totally occluded by the surface $s_1$. \(\square\)

Intuition 22 We cannot have interpenetrating surfaces.

The scenes in Figure 7.2(c) and (d) intuitively occur with penetrating surfaces that are mutually occluding, that is, there is a structure $\mathcal{M} \in \mathcal{L}_{\text{obscures}}$ such that
\[ \mathcal{M} \models (\exists s, s') \text{obscures}(s, s', t) \land \text{obscures}(s', s, t) \]

In this work, however, we will be concerned only with nonpenetrating surfaces. Thus, none of the following examples correspond to intuitive structures.

Example: Consider the scene in Figure 7.2(c). There exists a structure in $\mathcal{L}_{\text{obscures}}$ such that
\[ \{e_1, e_2, e_3, e_4, s_1, s_2\} \subseteq \mathcal{M} \text{ and} \]
\[ \langle s_2, s_1 \rangle, \langle s_1, s_2 \rangle \in \text{ontop.} \]
\[ \langle e_1, s_1 \rangle, \langle e_2, s_1 \rangle, \langle e_3, s_2 \rangle, \langle e_4, s_2 \rangle \in \text{part} \]

In this structure, $s_1$ and $s_2$ are mutually occluding; $e_1$ and $e_2$ are edges in $s_1$ that contain points that are occluded by $s_2$, while $e_3$ and $e_4$ are edges in $s_2$ that contain points that are occluded by $s_1$. In this example, the surfaces are interpenetrating, where two edges of $s_1$ poke through $s_2$ and two edges of $s_2$ poke through $s_1$. \(\square\)
For the next intuition, we need the notion of convex components for surfaces.

**Definition 7.1** Let $M$ be a structure in $L_{obscures}$. We will say that two positions $t_1, t_2$ are in the same convex component of a surface $s$ iff

$$\langle s, t_1 \rangle, \langle s, t_2 \rangle \in \text{interior}$$

and

$$\langle s, t_1, t_2 \rangle \in \text{noncrossing}$$

In other words, the two positions $t_1, t_2$ are in the same convex component of $s$ iff drawing a segment between $t_1$ and $t_2$ will not cross any edge of the surface.

**Intuition 23** Two surfaces can mutually occlude each other only if they have different convex components. The ordering over two surfaces in a scene must be the same for all positions in the same convex components of both surfaces.

There are two ways to think about this intuition and anomalous occlusion is the following. One aspect of the intuition says that positions in the same convex components of both surfaces must satisfy the same assignment of $obscures$ at those positions.

**Example:** Consider the scenes in Figure 7.2. There is a structure $M$ in $L_{obscures}$ such that $\{s_1, s_2\} \subset M$ and

$$\langle s_1, t_1 \rangle, \langle s_1, t_2 \rangle, \langle s_2, t_1 \rangle, \langle s_2, t_2 \rangle \in \text{interior}$$

$$\langle s_1, t_1, t_2 \rangle, \langle s_2, t_1, t_2 \rangle \in \text{noncrossing}$$

Thus, if

$$\langle s_1, s_2, t_1 \rangle \in \text{obscures}$$

then we must also have

$$\langle s_1, s_2, t_2 \rangle \in \text{obscures}$$

as in scene (a).

If

$$\langle s_2, s_1, t_1 \rangle \in \text{obscures}$$

then we must also have

$$\langle s_2, s_1, t_2 \rangle \in \text{obscures}$$

as in scene (b).

Notice that the scenes in Figure 7.2(c) and (d) violate this intuition. □

The other way to think about this intuition is that we can allow structures which satisfy

$$M \models (\exists s, s')\, \text{ontop}(s, s') \land \text{ontop}(s', s)$$

when the occluded points in the surfaces have disjoint sets of boundary edges or the positions are not noncrossing in one of the surfaces. In other words, mutually occluding surfaces are feasible if the positions that are obscured are in different convex components.
Example: Consider the scene in Figure 7.6(f). There exists a structure $\mathcal{M}$ in $\mathcal{L}_{\text{obsures}}$ such that 
\[ \{s_1, s_2, t_1, t_4\} \subseteq \mathcal{M} \] and
\[ (s_1, t_1), (s_1, t_4), (s_2, t_1), (s_2, t_4) \in \text{interior} \]
\[ (s_1, t_1, t_4) \in \text{noncrossing}, (s_2, t_1, t_4) \notin \text{noncrossing} \]
Thus we can have
\[ (s_1, s_2, t_1), (s_2, s_1, t_4) \in \text{obsures} \]
(as in scene (e)) or
\[ (s_2, s_1, t_1), (s_1, s_2, t_4) \in \text{obsures} \]
(as in scene (c)). □

Intuition 24 Whenever a surface is partially occluded, there is an occluding edge. There cannot be “patches” interior to the surface which are obscured without such an edge.

Example: Consider the scene in Figure 7.2(d). There exists a structure in $\mathcal{L}_{\text{obsures}}$ such that 
\[ \{e_1, e_2, e_3, e_4, p_1, p_2, s_1, s_2\} \subseteq \mathcal{M} \] and
\[ (s_1, s_2, t_1) \in \text{obsures} \]
\[ (e_1, s_2), (e_2, s_2) \in \text{part} \]
In Figure 7.2d, the shaded patch represents the extrusion of $s_1$ through $s_2$, containing the position $t_1$ where part of surface $s_1$ that $\text{obsures}$ a position interior to surface $s_2$. That is, part of $s_1$ penetrates $s_2$, and the shaded patch is what “pokes through” $s_2$. Intuitively, however, $e_1$ and $e_2$ should be occluding edges in $s_2$, and at all positions interior to $s_2$. □

Example: Consider the scene in Figure 7.1(a). There exists a structure $\mathcal{M} \in \mathcal{L}_{\text{obsures}}$ such that we have the following occluding edges:
\[ (e_1, s_2), (e_2, s_2), (e_1, s_3), (e_2, s_3), (e_3, s_3), (e_5, s_3), (e_8, s_3), (e_7, s_4), (e_8, s_4) \in \text{occluding} \]
In the scene Figure 7.1(b), there are no occluding edges. □

7.1.2 Structures for Obscured Scenes

Before characterizing the class of intended structures for these relations, it is useful to define the following substructures.

Definition 7.2 Given a structure $\mathcal{M}$ in $\mathcal{L}_{\text{interior}}$, a local surface ordering is a substructure $\langle S, t, < \rangle$ with the following properties:
Figure 7.2: Ambiguity in the definition of obscurers.
• \( S \) is a set of surfaces in \( M \);
• For some surface \( s \in M \), we have \( s \in S \) iff
  \[ (s, t) \in \text{interior} \]
• \( < \) is a discrete partial ordering;
• There exists a unique minimal element in \( S \) with respect to \( < \).

Local surface orderings formalize Intuitions 19 and 20. The intuition that surfaces can only be obscured at positions interior to both surfaces is captured by the condition in the definition of local surface orderings which restricts the set of surfaces to those which are interior at the position of the ordering. The ordering itself captures the intuition of the “depth” of a surface in the scene, and the minimal surface in the ordering corresponds to the surface which is intuitively on top of all other surfaces at some position.

Consider the scene in Figure 7.1(a). The local surface ordering at \( t_1 \) is

\[ s_1 < s_2 \]

The local surface ordering at \( t_2 \) is

\[ s_1 < s_2, s_2 < s_3, s_1 < s_3 \]

The local surface ordering at \( t_3 \) is

\[ s_1 < s_3 \]

The local surface ordering at \( t_4 \) is

\[ s_2 < s_3, s_2 < s_4 \]

However, the surfaces \( s_3 \) and \( s_4 \) are incomparable in the ordering, since they are abutting each other at the position \( t_4 \).

**Definition 7.3** Two local surface orderings \( (S_1, t_1, <^{S_1}) \) and \( (S_2, t_2, <^{S_2}) \) are order isomorphic with respect to two surfaces \( s_1, s_2 \) iff

\[ s_1 <^{S_1} s_2 \iff s_1 <^{S_2} s_2 \]

In the preceding example with the scene in Figure 7.1(a), the local surface orderings at the positions \( t_1 \) and \( t_2 \) are order-isomorphic, and the local surface orderings at the positions \( t_2 \) and \( t_3 \) are order-isomorphic.

**Definition 7.4** Given a structure \( M \) in \( L_{\text{interior}} \), a global surface ordering is a substructure which is the set of local surface orderings for each position in \( M \) such that the local surface orderings for positions which are in the same convex components of two surfaces must be order isomorphic with respect to these two surfaces.

Global surface orderings correspond to Intuition 23 – positions which are in the same convex components of two surfaces in a scene must agree on the obscures ordering for the two surfaces. Thus,
mutually obscuring surfaces are only allowed if the positions at which the surfaces are obscured are in different convex components of the surfaces. This also captures Intuition 22, since interpenetrating surfaces mutually occlude each other at positions which are in the same convex component.

**Definition 7.5** Given a structure $\mathcal{M} \in \mathcal{L}_{\text{obscures}}$, an overlap graph $G = (N, A)$ is a directed graph such that $N$ is the set of surfaces in $\mathcal{M}$, $G$ contains 2-cycles only if one of the surfaces has multiple distinct convex components, and $(s_1, s_2) \in A$ iff there exists a unique set of edges which are part of $s_1$ and which contain points $p_1$ such that $(p_1, t_1) \in \text{position and } s_1 < s_2$ in the local surface ordering at each position $t_1$.

The overlap graph for the scene in Figure 7.1(a) is shown in Figure 7.3. Note that for each arc in the graph, there exists an occluding edge.

We are now ready to define the class of structures $\mathcal{M}_i^{\text{obscures}}$ which we will later show to be equivalent to the models of $T_{\text{occ}}$:

**Definition 7.6** Let $\mathcal{M}_i^{\text{obscures}}$ be a set of structures in $\mathcal{L}_{\text{obscures}}$ with the following properties:

1. $\mathcal{M}$ is an extension of a structure in $\mathcal{M}_i^{\text{interior}}$;

2. The extensions of obscurses and abuts are isomorphic to a global surface ordering such that:
   
   (a) $\langle s_1, s_2, t \rangle \in \text{obscures}$
   
   iff $s_1 < s_2$ in the local surface ordering at $t$;

   (b) $\langle s_1, s_2, t \rangle \in \text{abuts}$
   
   iff $s_1$ and $s_2$ are incomparable in the local surface ordering at $t$.

3. The extension of partial occludes is isomorphic to an overlap graph $G = (N, A)$ such that
   
   if $\langle s_1, s_2 \rangle \in \text{partial occludes}$ then $s_1, s_2 \in N$ and $(s_1, s_2) \in A$.

4. Suppose a global surface ordering $\langle S, < \rangle$ and an overlap graph $G = (N, A)$ are substructures of $\mathcal{M}$. Then $(s_1, s_2) \in A$ iff the local surface orderings for a proper subset of positions interior to $s_2$ are order isomorphic to a proper subset of positions which are interior to $s_1$.

5. The extension of total occludes is isomorphic to a quasi-order $\prec$ such that $\langle s_1, s_2 \rangle \in \text{total occludes}$ iff $s_1 \prec s_2$. 

---

**Figure 7.3:** Overlap graph for the scene in Figure 7.1(a).
6. Suppose a global surface ordering \((S, <)\) and the total occludes quasi-order are substructures of \(\mathcal{M}\). Then \(s_1 < s_2\) iff the local surface orderings for all positions interior to \(s_2\) are order isomorphic to the positions interior to \(s_1\).

7. The extension of ontop is isomorphic to the graph which is the sum of the graphs for partial occludes and total occludes.

8. Let \(\pi(e)\) be a function which assigns an edge to the surface containing the edge.

\[
\langle e, s \rangle \in \text{occluding}
\]

iff there exists an overlap graph \(G = (N, A)\) in \(\mathcal{M}\) such that

\[
(\pi(e), s) \in A
\]

Each of the conditions in the definition of \(\mathcal{M}_1^{\text{obscures}}\) characterizes one of the following properties:

- the extension of a relation denoted by a predicate symbol in \(L_{\text{obscures}}\); in this case, we provide a structure which is isomorphic to the extension of the relation:

- the relationship among the substructures which are isomorphic to the extension of one of the relations:

Condition (2) characterizes the extensions of obscures and abuts, and corresponds to Intuitions 19 and 19. Condition (3) characterizes the extension of partial occludes, while condition (4) constrains the relationship between the extension of obscures and the extension of partial occludes. Condition (5) characterizes the extension of total occludes, while condition (6) constrains the relationship between the extension of obscures and the extension of total occludes. Together, these four conditions correspond to Intuition 21. Condition (7) characterizes the extension of ontop with respect to partial occludes and total occludes. Condition (8) characterizes the extension of occluding (i.e. occluding edges).

Note that we are not considering surfaces which are self-occluding. In such a case, obscures would not be isomorphic to a local surface ordering, since we would have \(\langle s, s, t \rangle \in \text{obscures}\).

Theorem 7.1 Existence Theorem for \(\mathcal{M}_1^{\text{obscures}}\)

The class of structures \(\mathcal{M}_1^{\text{obscures}}\) exists and is nonempty.

Proof: We first need to show that the following classes of structures exist:

1. local surface orderings
2. global surface orderings
3. overlap graphs

Since local surface orderings are a class of partial orderings, it is obvious that they exist.

Global surface orderings are the union of order-isomorphic local surface orderings. Since every surface has at least one convex component, we are guaranteed the existence of a set of order-isomorphic local surface orderings, and hence we are also guaranteed the existence of global surface orderings.

Overlap graphs are directed graphs; the only restriction is that the graphs can contain 2-cycles only if the surfaces have multiple convex components. For surfaces with unique convex components, overlap graphs are arbitrary directed graphs, and hence they exist.
In addition to guaranteeing the existence of these substructures, conditions (4) and (5) in the definition of $\mathcal{M}_i^{\text{obsures}}$ place restrictions on the combination of the substructures. We therefore need to show that these combinations of the substructures exist.

To satisfy condition (4), we need to show that given any global surface ordering, we are guaranteed the existence of an overlap graph. Since the arcs in the graph exist only if the positions interior to one surface are order-isomorphic to a set of positions interior to the other surface, the definition of global surface orderings guarantees that there will be 2-cycles in the graph only if there exist distinct sets of positions which are in different convex components of the surfaces; positions in the same convex component must be order-isomorphic.

To satisfy condition (5), we need to show that given any global surface ordering, we are guaranteed the existence of a relation which is a quasi-order over the surfaces. Since local surface orderings are substructures of the global surface ordering, and all positions interior to one surface are order-isomorphic to the positions interior to the other surface, the transitivity and irreflexivity of the relation follow from the definition of local surface ordering, which is a partial ordering. □

### 7.1.3 Axiomatizing $\text{obsures}$

We now introduce the axioms in the CardWorld module $T_{\text{obsures}}$ (Figure 7.4 and Figure 7.5) whose models will be structures in $\mathcal{M}_i^{\text{obsures}}$. A scene object is occluded when there is a surface ontop of the object: the relation $\text{ontop}$ is used to capture this relationship (axiom 7.11 of $T_{\text{obsures}}$). However, it is insufficient to define the notion of occlusion, since it is possible for surfaces to be mutually occluding. We therefore need to say that one surface occludes another at a particular position. For this reason we define $\text{ontop}$ in terms of another relation, $\text{obsures}$. This relation formalizes the notion of one surface occluding another surface at a position; a surface is $\text{ontop}$ if it $\text{obsures}$ the surface at some position.

Axiom 7.1 of $T_{\text{obsures}}$ is the sentence that specifies the relationship between occlusion and the interior of a surface. If a position $t$ is $\text{interior}$ to two surfaces, then one surface must obscure the other surface at that position, or the surfaces must be abutting at that position. The interesting aspect of this axiom is the ambiguity concerning which surface is obscuring the given position. That is, given a position $t$ alone, if it is interior to two surfaces, we cannot determine which surface is obscured and which is obscuring, or if they are abutting.

Most of the axioms of $T_{\text{obsures}}$ (in particular, axioms 7.1 - 7.7 and axiom 7.10) are the constraints on local surface orderings, which place constraints on the extension of $\text{obsures}$ at some fixed position. These axioms specify that the extension of $\text{obsures}$ is isomorphic to a partial ordering, and that abutting surfaces are incomparable in this ordering. Axiom 7.10 states that the occluding surface be the minimal in the $\text{obsures}$ ordering at the position of the point being occluded. This prevents an infinite ascending stack of occluding surfaces in any model, which would cause problems for any theory of occlusion. For example, an undepicted point is obscured by a surface; the points in this surface must also either be depicted or obscured by some surface. If we allowed infinite stacks of surfaces, we could have undepicted occluded points but none of the occluding surfaces would be depicted; this violates the intuitions we have about the nature of occlusion. Axiom 7.10 can be considered to be minimal obscuring surface
If a position is interior to two different surfaces, then either one of the surfaces obscures the other at that position, or the surfaces are abutting each other at that position.

\[(\forall s_1, s_2, t) \text{interior}(s_1, t) \land \text{interior}(s_2, t) \equiv \]

\[(\text{obscures}(s_1, s_2, t) \lor \text{obscures}(s_2, s_1, t) \lor (\exists e_1, e_2) \text{part}(e_1, s) \land \text{part}(e_2, s_2) \land \text{abuts}(e_1, e_2, t)) \quad (7.1)\]

Surfaces cannot obscure each other at the same position.

\[(\forall s_1, s_2, t) \text{obscures}(s_1, s_2, t) \supset \lnot \text{obscures}(s_2, s_1, t) \]

(7.2)

At any position, the obscures relation is transitive.

\[(\forall s_1, s_2, s_3, t) \text{obscures}(s_1, s_2, t) \land \text{obscures}(s_2, s_3, t) \supset \text{obscures}(s_1, s_3, t) \quad (7.3)\]

Surfaces which are abutting each other at a position cannot obscure each other at that position.

\[(\forall e_1, s_1, e_2, s_2, t) \text{abuts}(e_1, e_2, t) \land \text{part}(e_1, s_1) \land \text{part}(e_2, s_2) \supset \lnot \text{obscures}(s_1, s_2, t) \quad (7.4)\]

If one surface obscures a surface which is abutting against another surface at the same position, then the first surface obscures both abutting surfaces at that position.

\[(\forall e_1, e_2, s_1, s_2, s_3, t) \text{obscures}(s_1, s_2, t) \land \text{abuts}(e_1, e_2, t) \land \text{part}(e_1, s_1) \land \text{part}(e_2, s_2) \supset \text{obscures}(s_1, s_3, t) \quad (7.5)\]

At any position, the relation abuts is symmetric in its edge arguments.

\[(\forall e_1, e_2, t) \text{abuts}(e_1, e_2, t) \supset \text{abuts}(e_2, e_1, t) \quad (7.6)\]

Abutting edges must be parts of distinct surfaces.

\[(\forall e_1, e_2, s, t) \text{abuts}(e_1, e_2, t) \land \text{part}(e_1, s) \supset \lnot \text{part}(e_2, s) \quad (7.7)\]

The extension of obscures is the same for all positions in the same convex components of the surfaces.

\[(\forall s_1, s_2, t_1, t_2) \text{obscures}(s_1, s_2, t_1) \land \text{interior}(s_1, t_2) \land \text{interior}(s_2, t_2) \land \text{noncrossing}(s_1, t_1, t_2) \land \text{noncrossing}(s_2, t_1, t_2) \supset \text{obscures}(s_1, s_2, t_2) \quad (7.8)\]

The obscures ordering is discrete.

\[(\forall s_1, s_2, t) \supset (\exists s_3) \text{obscures}(s_1, s_3, t) \land (\lnot (\exists s_4) \supset \text{obscures}(s_1, s_4, t) \land \text{obscures}(s_4, s_3, t)) \quad (7.9)\]

At any position, there exists a minimal surface which is not obscured by any other surface.

\[(\forall s_1, t) \supset (\exists s_2) (\neg \exists s_3) \supset \text{interior}(s_2, t) \land \lnot \text{obscures}(s_3, s_2, t) \quad (7.10)\]

Figure 7.4: T_{obscures}: Axioms for obscuring surfaces.
A surface is ontop of another surface iff there exists a position at which the surface obscures it.

\[(\forall s_1, s_2) \text{ontop}(s_1, s_2) \equiv (\exists t) \text{obsures}(s_1, s_2, t)\] (7.11)

One surface partially occludes another surface iff there exists a position at which the first surface obscures the second surface and another position interior to the second surface at which it is not obscured.

\[(\forall s, s') \text{partial\_occludes}(s, s') \equiv (\exists t_1, t_2) \text{obsures}(s, s', t_1) \land \neg \text{obsures}(s, s', t_2) \land \text{interior}(s', t_2)\] (7.12)

One surface totally occludes another surface iff the first surface obscures the second surface at all positions interior to the second surface.

\[(\forall s, s') \text{total\_occludes}(s, s') \equiv (\exists t_1) \text{obsures}(s, s', t_1) \land ((\forall t_2) \text{interior}(s', t_2) \supset \text{obsures}(s, s', t_2))\] (7.13)

An edge is an occluding edge iff there exists a point in the edge, and the surface containing the edge obscures another surface at the position of that point.

\[(\forall e, s, t) \text{occluding}(e, s) \equiv \]

\[(\exists p, t, s') \text{part}(e, s') \land \text{part}(p, e) \land \text{position}(p, t) \land \text{obsures}(s', s, t)\] (7.14)

---

Figure 7.5: $T_{\text{obsures}}$: Conservative definitions.
assumption since it asserts the existence of a topmost surface that is minimal in the $obscures$ ordering at any position.

Axiom 7.8 specifies the constraint on global surface orderings, and is perhaps one of the most important axioms within $T_{obscures}$. The problem is to consistently combine the local surface orderings for all positions which are interior to a set of surfaces. If this is not done properly, we can get some very strange and unintuitive results.

In the earlier examples that motivated Intuitions 22 and 23, there were cases of anomalous occlusion and interpenetrating surfaces. Axiom 7.8 of $T_{obscures}$ eliminates these cases. We can illustrate axiom 7.8 using the original examples of anomalous occlusion. In Figure 7.2(e), the shaded portion is the set of positions that are interior to both surfaces. Suppose we have a model $\mathcal{M}$ of $T_{scene} \cup T_{fg} \cup T_{interior} \cup T_{obscures}$ such that

$$\langle s_1, s_2, t_1 \rangle \in obscures$$

By axiom 7.8 of $T_{obscures}$, all models of $T_{scene} \cup T_{fg} \cup T_{interior} \cup T_{obscures}$ agree on the extension of $obscures$ at the position $t_2$ since the two positions are in the same convex component in both surfaces. Thus if

$$\langle s_1, t_1, t_2 \rangle \in noncrossing$$

then we must have

$$\langle s_1, s_2, t_2 \rangle \in obscures$$

In this way, the scenes in Figure 7.2(c) and (d) are eliminated by axiom 7.8 of $T_{obscures}$, since in these scenes, some of the positions in the shaded portion of Figure 7.2(e) satisfy

$$\langle s_1, s_2, t_1 \rangle \in obscures$$

and some of the positions in the shaded portions of Figure 7.2(e) satisfy

$$\langle s_2, s_1, t_2 \rangle \in obscures$$

The scenes in Figure 7.2(a) and (b) are the only intuitive possibilities consistent with axiom 7.8 of $T_{obscures}$.

**Example:** Consider Figure 7.6. The diagram in (b) displays the scene with all hidden edges visible, and scenes (a), (c), and (e) are feasible scenes. Using $T_{obscures}$, we can see why these scenes are feasible.

Consider the scene in Figure 7.6a. Let $\mathcal{M}$ be a model of $T_{scene} \cup T_{fg} \cup T_{interior} \cup T_{obscures}$ such that

$$\{e_1, e_2, e_3, e_4, e_5, e_6, e_7, e_8, s_1, s_2\} \subset M.$$
Figure 7.6: Eliminating anomalous occlusion.
In Figure 7.6d, the two darkly shaded portions are the set of positions that are interior to both surfaces. Axiom 7.8 is equivalent to saying that either all positions in one shaded portion must satisfy

\[ (s_1, s_2, t) \in \text{obscures} \]

or all positions in one shaded portion must satisfy

\[ (s_2, s_1, t) \in \text{obscures} \]

Note that since positions in the two shaded portions are not noncrossing with respect to each other, axiom 7.8 does not constrain the assignment of obscures between positions in different shaded portions.

Consider Figure 7.6(f). Suppose that

\[ (s_1, s_2, t_1) \in \text{obscures} \]

We have

\[ (e_6, s_2), (e_3, s_1) \in \text{part} \]
\[ (e_6, t_2), (e_8, t_5), (e_3, t_1), (e_3, t_2) \in \text{boundary} \]
\[ (s_1, t_1, t_5), (s_1, t_2, t_5), (s_2, t_1, t_5), (s_2, t_2, t_5) \in \text{noncrossing} \]

By axiom 7.8 of \( T_{\text{obscures}} \), we get

\[ (s_1, s_2, t_2), (s_1, s_2, t_5) \in \text{obscures} \]

This is the case in Figure 7.6(a) and (e).

Suppose that

\[ (s_2, s_1, t_1) \in \text{obscures} \]

By axiom 7.8 of \( T_{\text{obscures}} \), we get

\[ (s_2, s_1, t_2), (s_2, s_1, t_5) \in \text{obscures} \]

This is the case in Figure 7.6(c).

For the other set of positions that are interior to both surfaces, we reason in a similar way. Suppose that

\[ (s_1, s_2, t_1) \in \text{obscures} \]

We have

\[ (e_1, s_1), (e_5, s_2) \in \text{part} \]
\[ (e_1, t_3), (e_5, t_4) \in \text{boundary} \]
The remaining relations (ontop, partial\_occludes, total\_occludes, and occluding) can be given conservative definitions: these are found in Figure 7.5. An example of these relations can be found in the scenes of Figure 7.1.

**Example:** Consider the scene in Figure 7.1(a). There exists a structure $\mathcal{M} \in \mathcal{L}_{obscures}$ such that

$$\{s_1, s_2, s_3, s_4, e_1, e_2, e_3, e_4, e_5, e_6, e_7, e_8\} \subseteq M$$

and which has the following properties.

Since there exist positions at which $s_1$ obscures $s_2$, as well as positions which are interior to $s_2$ but not interior to $s_1$, the definition gives us

$$\langle s_1, s_2 \rangle \in \text{partial\_occludes}$$

Similarly, we also get

$$\langle s_1, s_3 \rangle, \langle s_2, s_4 \rangle \in \text{partial\_occludes}$$

The definition of occluding gives us

$$\langle e_1, s_2 \rangle, \langle e_2, s_2 \rangle, \langle e_2, s_3 \rangle, \langle e_3, s_3 \rangle, \langle e_8, s_4 \rangle, \langle e_8, s_4 \rangle, \langle e_7, s_4 \rangle \in \text{occluding}$$

Notice that for every partially occluded surface, there exists at least one occluding edge. We will later prove that this is always the case.

**Example:** Consider the scene in Figure 7.1(b). There exists a structure $\mathcal{M} \in \mathcal{L}_{obscures}$ such that

$$\{s_1, s_2\} \subseteq M$$

and which has the following properties.

$$\langle s_1, s_2 \rangle \in \text{total\_occludes}$$

since $s_1$ obscures $s_2$ at all positions interior to $s_2$. Note that in this case of total occlusion, there are no occluding edges.
7.1.4 Satisfiability Theorem for Obscuring Surfaces

We can now show that $T_{\text{obscures}}$ is consistent, and that the structures in $M_i^{\text{obscures}}$ are models of $T_{\text{obscures}}$.

**Theorem 7.2** Any structure in $M_i^{\text{obscures}}$ is a model of $T_{\text{scene}} \cup T_{\text{global}} \cup T_{\text{interior}} \cup T_{\text{interior}} \cup T_{\text{obscures}}$.

**Proof:** Let $M$ be a structure in $M_i^{\text{obscures}}$.

$$M \models (\forall s_1, s_2, t) \lor \text{interior}(s_1, t) \land \text{interior}(s_2, t) \equiv$$

$$(\exists e_1, e_2) \lor \text{part}(e_1, s) \land \text{part}(e_2, s) \land \text{abuts}(e_1, e_2, t))$$

iff for any variable assignment $\sigma$, for all elements $\sigma(s_1), \sigma(s_2), \sigma(t) \in M$, we have

$$\langle \sigma(s_1), \sigma(t) \rangle \in \text{interior and } \langle \sigma(s_2), \sigma(t) \rangle \in \text{interior}$$

iff one of the following holds:

- $\langle \sigma(s_1), \sigma(s_2), \sigma(t) \rangle \in \text{obscures}$
- $\langle \sigma(s_2), \sigma(s_1), \sigma(t) \rangle \in \text{obscures}$
- There exist edges $e_1, e_2 \in M$ such that

$$\langle e_1, e_2, \sigma(t) \rangle \in \text{abuts}$$

This follows because $\text{obscures}$ is isomorphic to a partial ordering with minimal elements, and $\text{abuts}$ is isomorphic to incomparability among elements in the ordering.

$$M \models (\forall s_1, s_2, t) \lor \text{obscures}(s_1, s_2, t) \lor \neg \text{obscures}(s_2, s_1, t)$$

iff for any variable assignment $\sigma$, for all elements $\sigma(s_1), \sigma(s_2), \sigma(t) \in M$, we have

$$\langle \sigma(s_1), \sigma(s_2), \sigma(t) \rangle \in \text{obscures} \text{ iff } \langle \sigma(s_2), \sigma(s_1), \sigma(t) \rangle \notin \text{obscures}.$$ 

This follows because the extension of $\text{obscures}$ at every position is isomorphic to a partial ordering, which is irreflexive.

$$M \models (\forall s_1, s_2, s_3, t) \lor \text{obscures}(s_1, s_2, t) \lor \text{obscures}(s_2, s_3, t) \lor \text{obscures}(s_1, s_3, t)$$

iff for any variable assignment $\sigma$, for all elements $\sigma(s_1), \sigma(s_2), \sigma(s_3), \sigma(t) \in M$, if

$$\langle \sigma(s_1), \sigma(s_2), \sigma(t) \rangle \in \text{obscures and } \langle \sigma(s_2), \sigma(s_3), \sigma(t) \rangle \in \text{obscures},$$

then we have

$$\langle \sigma(s_1), \sigma(s_3), \sigma(t) \rangle \in \text{obscures and}$$

This follows because the extension of $\text{obscures}$ at every position is isomorphic to a partial ordering, which is transitive.

$$M \models (\forall e_1, e_2, t) \lor \text{abuts}(e_1, e_2, t) \lor \text{abuts}(e_2, e_1, t)$$

iff for any variable assignment $\sigma$, for all elements $\sigma(e_1), \sigma(e_2), \sigma(t) \in M$, we have

$$\langle \sigma(e_1), \sigma(e_2), \sigma(t) \rangle \in \text{abuts} \text{ iff } \langle \sigma(e_2), \sigma(e_1), \sigma(t) \rangle \in \text{abuts}.$$ This follows because the edges in $\text{abuts}$ are incomparable in the local surface ordering, so that the relation is symmetric with respect to edges.

$$M \models (\forall s_1, s_2, t_1, t_2) \lor \text{obscures}(s_1, s_2, t) \lor \text{interior}(s_1, t_2) \lor \text{interior}(s_2, t_2) \lor \text{noncrossing}(s_1, t_1, t_2) \lor \text{noncrossing}(s_2, t_1, t_2)$$
\( \mathcal{M} \models (\forall s_1, t) \text{interior}(s_1, t) \supset (\exists s_2)(\forall s_3) \text{interior}(s_2, t) \land \neg \text{obscures}(s_3, s_2, t) \)

iff for any variable assignment \( \sigma \), for all elements \( \sigma(s_1), \sigma(s_2), \sigma(t) \in M \) there exists an element \( \sigma(s_2) \in M \) such that for all elements \( \sigma(s_3) \in M \) we have

\[ \langle \sigma(s_2), \sigma(t) \rangle \in \text{interior} \text{ and } \sigma(s_2), \sigma(s_3), \sigma(t) \notin \text{obscures}. \]

This is guaranteed because the partial ordering isomorphic to \text{obscures} has a unique initial surface at each position.

\( \mathcal{M} \models (\forall s_1, s_2, t) \text{obscures}(s_1, s_2, t) \supset (\exists s_3)\text{obscures}(s_1, s_3, t) \land (\neg (\exists s_4)\text{obscures}(s_1, s_4, t) \land \text{obscures}(s_4, s_2, t)) \)

iff for any variable assignment \( \sigma \), for any surfaces \( \sigma(s_1), \sigma(s_2) \) and position \( \sigma(t) \), if

\[ \langle \sigma(s_1), \sigma(s_2), \sigma(t) \rangle \in \text{obscures} \]

then there exists a surface \( \sigma(s_3) \) such that

\[ \langle \sigma(s_1), \sigma(s_3), \sigma(t) \rangle \in \text{obscures} \]

and there does not exist any intervening surface between \( \sigma(s_1) \) and \( \sigma(s_2) \). This holds because the partial ordering isomorphic to \text{obscures} is discrete.

Now consider the definition of \text{partial\_occludes}: we need to show that the extension of \text{partial\_occludes} is characterized by the definition in \text{Tobscures}.

For any variable assignment \( \sigma \),

\( \mathcal{M}, \sigma \models \text{partial\_occludes}(s_1, s_2) \)

iff there exist positions \( \sigma(t_1), \sigma(t_2) \in M \) such that

\[ \langle \sigma(s_1), \sigma(s_2), t_1 \rangle \in \text{obscures} \text{ and } \langle \sigma(s_2), \sigma(t_2) \rangle \in \text{interior}, \]

but

\[ \langle \sigma(s_1), \sigma(s_2), \sigma(t_2) \rangle \notin \text{obscures} \]

This follows from condition (3) in the definition of \( \mathcal{M}_1^\text{obscures} \), which states that the local surface orderings for a proper subset of positions interior to \( \sigma(s_2) \) are order isomorphic to a proper subset of positions which are interior to \( \sigma(s_2) \).

Now consider the definition of \text{total\_occludes}: we need to show that the extension of \text{total\_occludes} is characterized by the definition in \text{Tobscures}.

For any variable assignment \( \sigma \),

\( \mathcal{M}, \sigma \models \text{total\_occludes}(s_1, s_2) \)

iff for all positions \( \sigma(t_1), \sigma(t_2) \in M \), if \( \langle \sigma(s_1), \sigma(s_2), t_1 \rangle \in \text{obscures} \text{ and } \langle \sigma(s_2), \sigma(t_2) \rangle \in \text{interior} \), then

\[ \langle \sigma(s_1), \sigma(s_2), \sigma(t_2) \rangle \in \text{obscures} \]

This follows from condition (5) in the definition of \( \mathcal{M}_1^\text{obscures} \), which states that all local surface orderings for positions interior to \( \sigma(s_2) \) are order isomorphic to \( \sigma(s_2) \). \( \square \)
7.1.5 Existence of Occluding Edges

One of the primary intuitions used to motivate the axioms for \textit{obsures} was that it is the interior or figure of a surface that is responsible for occlusion, and that boundary edges play an important role in defining how the surface occludes other scene objects. This role is encapsulated in axiom 7.14 of \textit{Tobsures}, which is the definition of occluding edges. An edge \(e\) occludes a surface \(s'\) iff \(e\) is part of a surface \(s\) that obscures \(s'\) at some position \(t\) and which has a point \(p\) at that position.

The notion of occluding edges is closely linked to the intuitions that we use to eliminate the unintuitive scenes earlier in the chapter. In the cases of anomalous occlusion in Figures 7.2(c) and (d), occlusion is done within the figure of the surface: that is, there are occluded parts in the obscured surface whose positions are intuitively within the obscuring surface, yet there are no occluding edges.

Example: Consider the scene in Figure 7.2(d). There is a structure \(\mathcal{M}\) in \(L_{\text{obsures}}\) such that

\[
(s_1, s_2, t_1), (s_2, s_1, t_2) \in \text{obsures}
\]

for positions \(t_1, t_2\), yet there is no edge in either surface that is responsible for the occlusion at either of these positions. Thus, this structure cannot be in \(\mathcal{M}_{\text{obsures}}^*\).

We therefore need to guarantee that occlusion is always done at the boundary of the surface. The following lemma guarantees the existence of occluding edges whenever one surface \textit{obsures} another surface at some position.

**Lemma 7.1 Occluding Edge Existence Lemma**

\[
T_{\text{scene}} \cup T_{fg} \cup T_{\text{interior}} \cup T_{\text{obsures}} \models (\forall s, s', t) \text{obsures}(s, s', t) \land \text{partial-occludes}(s, s') \supset (\exists e) \text{occluding}(e, s') \land \text{part}(e, s)
\]

**Proof:** Suppose

\[
T_{\text{scene}} \cup T_{fg} \cup T_{\text{interior}} \cup T_{\text{obsures}} \models (\exists s, s', t) \text{obsures}(s, s', t) \land \text{partial-occludes}(s, s')
\]

By axiom 7.1 of \(T_{\text{obsures}}\) and the definition of \textit{partial-occludes}, we must have

\[
T_{\text{scene}} \cup T_{fg} \cup T_{\text{interior}} \cup T_{\text{obsures}} \models (\forall s, s', t) \text{obsures}(s, s', t) \land \text{partial-occludes}(s, s', t) \supset (\exists t_2) \text{interior}(s, t) \land \text{interior}(s', t) \land \text{interior}(s, t_2) \land \neg \text{interior}(s', t_2)
\]

By axiom 6.7, we know that the position of one of the points in one of the boundary edges is interior to both surfaces:

\[
T_{\text{scene}} \cup T_{fg} \cup T_{\text{interior}} \cup T_{\text{obsures}} \models (\forall s, s', t) \text{obsures}(s, s', t) \land \text{partial-occludes}(s, s', t) \supset (\exists e, p, t_2) \text{interior}(s, t) \land \text{interior}(s', t) \land \text{boundary}(e, t) \land \text{part}(e, s) \land \text{part}(p, e) \land \text{position}(p, t_2) \land \text{interior}(s, t_2)
\]

Axiom 7.8 of \(T_{\text{obsures}}\) gives us
By the definition of \textit{occluding} we get

\[
T_{\text{scene}} \cup T_{\text{fg}} \cup T_{\text{interior}} \cup T_{\text{obscures}} \models (\forall s, s', e, p, t, t_2) \text{obscures}(s, s', t) \land \text{partial-occludes}(s, s', t) \\
\land \text{interior}(s, t) \land \text{interior}(s', t) \\
\land \text{boundary}(e, t) \land \text{part}(e, s) \land \text{part}(p, e) \land \text{position}(p, t_2) \land \text{interior}(s, t_2) \\
\supset \text{obscures}(s, s', t_2)
\]

Thus we are guaranteed the existence of an occluding edge that is part of the obscuring surface.

\(\Box\)

This lemma can be considered to be an alternative characterization of occlusion, since it requires that occlusion always take place at a boundary edge of a surface. Axiom 7.8 plays a crucial role: it is the axiom that forces the boundary edges for an occluding surface to be unobscured. In a sense, the occluding edge provides the assignment of \textit{obscures} literals for all positions for which it is the occluding edge.

\textbf{Example:} Consider the scene in Figure 7.6(c). There exists a model \(\mathcal{M}\) of \(T_{\text{obscures}} \cup T_{\text{interior}} \cup T_{\text{fg}} \cup T_{\text{scene}}\) with domain \(\{e_1, e_2, e_3, e_4, s_1, s_2\}\) such that

\[
(e_1, s_1), (e_2, s_2), (e_3, s_1), (e_4, s_2) \in \text{part} \\
(e_2, s_1), (e_3, s_2) \in \text{occluding}
\]

so that for all positions \(t_i\) in the upper shaded area of Figure 7.6(d) we have

\[
(s_2, s_1, t_i) \in \text{obscures}
\]

and for all positions \(t_j\) in the lower shaded area of Figure 7.6(d) we have

\[
(s_1, s_2, t_j) \in \text{obscures}
\]

In particular, using the positions in Figure 7.6(f), we get

\[
(s_2, s_1, t_1), (s_2, s_1, t_2), (s_1, s_2, t_3), (s_1, s_2, t_4) \in \text{obscures}
\]

In Figure 7.6(e), there exists a model \(\mathcal{M}\) of \(T_{\text{obscures}} \cup T_{\text{interior}} \cup T_{\text{fg}} \cup T_{\text{scene}}\) with domain
\{e_1, e_2, e_3, e_4, s_1, s_2\} such that

\(\langle e_1, s_1 \rangle, \langle e_2, s_2 \rangle, \langle e_3, s_3 \rangle, \langle e_4, s_2 \rangle \in \text{part}\)

\(\langle e_1, s_2 \rangle, \langle e_4, s_1 \rangle \in \text{occluding}\)

so that for all positions \(t_i\) in the upper shaded area of Figure 7.6(d) we have

\(\langle s_1, s_2, t_i \rangle \in \text{obsures}\)

and for all positions \(t_j\) in the lower shaded area of Figure 7.6(d) we have

\(\langle s_2, s_1, t_j \rangle \in \text{obsures}\)

In particular, using the positions in Figure 7.6(f), we get

\(\langle s_1, s_2, t_1 \rangle, \langle s_1, s_2, t_2 \rangle, \langle s_2, s_1, t_3 \rangle, \langle s_2, s_1, t_4 \rangle \in \text{obsures}\)

\(\square\)

In each of these examples, we can completely determine the extension of \(\text{obsures}\) by specifying the occluding edges.

Also note that the Occluding Edge Existence Lemma is independent of depiction, since it must be satisfied even if the occluding edges are themselves undepicted. Later in this chapter, we will consider theories in which the occluding edges are depicted, and explore the implications of this property for occlusion.

### 7.1.6 Axiomatizability Theorem for \(T_{\text{obsures}}\)

**Theorem 7.3** Let \(\mathcal{M}\) be a structure in \(L_{\text{obsures}}\) with the following properties:

- \(\mathcal{M}\) contains a substructure isomorphic to a structure in \(M^\text{polygon}_{i}\);
- \(\mathcal{M}\) contains a substructure isomorphic to a structure in \(M^\text{plane}_{i}\);
- \(\mathcal{M}\) is a model of \(T_{\text{scene}} \cup T_{\text{fg}} \cup T_{\text{interior}} \cup T_{\text{obsures}}\);

then \(\mathcal{M}\) is isomorphic to a structure in \(M^\text{obsures}_{i}\).

**Proof:** Let \(\mathcal{M}\) be a model of \(T_{\text{scene}} \cup T_{\text{fg}} \cup T_{\text{interior}} \cup T_{\text{obsures}}\) that satisfies the conditions in the hypothesis of the theorem.

We will first construct a graph which satisfies axioms 7.1 - 7.10, and then show that this graph is isomorphic to a local surface ordering.

Let \(G(t) = (N, A)\) be a graph such that

\[N = \{s : (s, t) \in \text{interior}\}\]

and

\[(s_1, s_2) \in A \Leftrightarrow (s_1, s_2, t) \in \text{obsures}\]
By axiom 7.1, $N$ is nonempty (it will contain at least one surface. since every surface has a nonempty set of positions interior to the surface). By axiom 7.2, there are no 1-cycles or 2-cycles in $G(t)$. By axiom 7.3, if $(s_1, s_2) \in A$ and $(s_2, s_3) \in A$, then we have $(s_1, s_3) \in A$.

By axioms 7.1 and 7.6, if $s_1, s_2 \in N$ but

$$(s_1, s_2, t) \notin \text{obscures}$$

then there must exist edges $e_1, e_2$ such that

$$(e_1, e_2, t), (e_2, e_1, t) \in \text{abuts}$$

By axiom 7.4, we must have $(s_1, s_2) \notin A$, so that $s_1$ and $s_2$ are incomparable.

By axiom 7.9, if $(s_1, s_2) \in A$, then there exists a surface $s_3$ such that $(s_1, s_3) \in A$, but there is no intervening surface between them.

By these four properties, $G(t)$ is a discrete partial ordering.

By axiom 7.10, there exists $s \in N$ such that for all $s' \in N$ we have $(s, s') \in A$, so that there exists a unique minimal element in $N$ with respect to the partial ordering.

Thus $G(t)$ is isomorphic to a local surface ordering.

By axiom 7.8, the union of $G(t)$ graphs for all positions will be isomorphic to a global surface ordering. Thus, condition (2) in the definition of $\mathcal{M}_t^{\text{obscures}}$ are satisfied by $\mathcal{M}$.

We next characterize the extension of total occludes within $\mathcal{M}$. To show that total occludes is a quasi-order, which follows from the next lemma:

**Lemma 7.2**

$$T_{\text{scene}} \cup T_{\text{fg}} \cup T_{\text{interior}} \cup T_{\text{obscures}} \models (\forall s_1, s_2) \text{ total occludes}(s_1, s_2) \supset \neg \text{total occludes}(s_2, s_1)$$

$$T_{\text{scene}} \cup T_{\text{fg}} \cup T_{\text{interior}} \cup T_{\text{obscures}} \models (\forall s_1, s_2, s_3) \text{ total occludes}(s_1, s_2) \land \text{total occludes}(s_1, s_3) \supset \text{total occludes}(s_1, s_3)$$

**Proof:** The first sentence follows from the definition of total occludes (axiom 7.13) and the irreflexivity of obscures (axiom 7.2).

The second sentence follows from the definition of total occludes (axiom 7.13) and the transitivity of obscures (axiom 7.3). $\Box$

Thus, condition (5) in the definition of $\mathcal{M}_t^{\text{obscures}}$ is satisfied by $\mathcal{M}$.

By Lemma 7.1, condition (8) in the definition of $\mathcal{M}_t^{\text{obscures}}$ are satisfied by $\mathcal{M}$.

The extension of partial occludes is characterized by Lemma 7.1, so that condition (3) in the definition of $\mathcal{M}_t^{\text{obscures}}$ is satisfied by $\mathcal{M}$.

For condition (4), for any variable assignment $\sigma$, by the definition of partial occludes, if

$$\mathcal{M}, \sigma \models \text{partial occludes}(s_1, s_2)$$

then there exist positions $\sigma(t_1), \sigma(t_2) \in M$, such that

$$\mathcal{M}, \sigma \models \text{obscures}(s_1, s_2, t_1) \land \text{interior}(s_2, t_2) \land \text{obscures}(s_1, s_2, t_2)$$

so that the local surface orderings at $\sigma(t_1)$ and $\sigma(t_2)$ are order isomorphic. However, by the definition of partial occludes there also exists a position $\sigma(t_3)$ such that

$$\mathcal{M}, \sigma \models \neg \text{obscures}(s_1, s_2, t_3) \land \neg \text{interior}(s_1, t_3) \land \text{interior}(s_2, t_3)$$
so that the set of positions at which \( \sigma(s_1) \) obscures \( \sigma(s_2) \) is a subset of the points which are interior to \( \sigma(s_2) \). Thus condition (4) in the definition of \( M_{i}^{obscures} \) is satisfied by \( M \).

For condition (6), for any variable assignment \( \sigma \), if
\[
M, \sigma \models total\text{-}occludes(s_1, s_2)
\]
then for all positions \( \sigma(t_1), \sigma(t_2) \in M \), if
\[
M, \sigma \models obscures(s_1, s_2, t_1) \wedge interior(s_2, t_2)
\]
then
\[
M, \sigma \models obscures(s_1, s_2, t_2)
\]
by the definition of \( total\text{-}occludes \), so that the local surface orderings at \( \sigma(t_1) \) and \( \sigma(t_2) \) are order isomorphic, and condition (6) in the definition of \( M_{i}^{obscures} \) is satisfied by \( M \).

Finally, we characterize the extension of \( ontop \) within \( M \). This follows from the next lemma:

**Lemma 7.3**

\[
T_{\text{scene}} \cup T_{fg} \cup T_{\text{interior}} \cup T_{\text{obscures}} \models (\forall s_1, s_2) ontop(s_1, s_2) \equiv total\text{-}occludes(s_1, s_2) \vee partial\text{-}occludes(s_1, s_2)
\]

**Proof:** One direction follows easily from the definitions:
\[
T_{\text{scene}} \cup T_{fg} \cup T_{\text{interior}} \cup T_{\text{obscures}} \models (\forall s_1, s_2) total\text{-}occludes(s_1, s_2) \supset ontop(s_1, s_2)
\]
\[
T_{\text{scene}} \cup T_{fg} \cup T_{\text{interior}} \cup T_{\text{obscures}} \models (\forall s_1, s_2) partial\text{-}occludes(s_1, s_2) \supset ontop(s_1, s_2)
\]

For the converse direction, we will show:
\[
T_{\text{scene}} \cup T_{fg} \cup T_{\text{interior}} \cup T_{\text{obscures}} \models
\\((\forall s_1, s_2) ontop(s_1, s_2) \wedge \neg partial\text{-}occludes(s_1, s_2) \supset total\text{-}occludes(s_1, s_2))
\]

By definition of \( partial\text{-}occludes \) and \( ontop \), we have:
\[
T_{\text{scene}} \cup T_{fg} \cup T_{\text{interior}} \cup T_{\text{obscures}} \models (\forall s_1, s_2, t) obscures(s_1, s_2, t) \wedge \neg partial\text{-}occludes(s_1, s_2) \supset total\text{-}occludes(s_1, s_2)
\]

\[
\supset [(\forall t_1, t_2) obscures(s_1, s_2, t_1) \supset (obscures(s_1, s_2, t_2) \vee \neg interior(s_2, t_2))]
\]

It is easy to see that the antecedent is equivalent to the definition of \( total\text{-}occludes \). \( \Box \)

Thus condition (7) in the definition of \( M_{i}^{obscures} \) is satisfied.

Since all of the conditions in the definition of \( M_{i}^{obscures} \) are satisfied, we have \( M \in M_{i}^{obscures} \). \( \blacksquare \)

As a simple corollary of this result, we also have the following characterization of the set of structures in \( M_{i}^{obscures} \) as a whole – given the set of local surface orderings at each position in one structure in \( M_{i}^{obscures} \), it determines the complete set of structures in \( M_{i}^{obscures} \).

**Corollary 7.1** Given a local surface ordering \( (S, t, \prec) \), for each surface \( s \in S \) there exists a structure in \( M_{i}^{obscures} \) such that \( s \) is the minimal surface in the ordering.
7.2 Occlusion and Depiction

We have so far characterized the scene relation *obscures*, but we have yet to specify the relationship between *obscures* and depiction, which is the central intuition for occlusion. In this section, we will introduce the final CardWorld module $T_{occ}$ which will axiomatize the intuitions for occlusion. The axioms in $T_{occ}$ (Figure 7.10) contain both the scene predicate *obscures* and the depiction predicate $\Delta$: these are occlusion axioms insofar as occlusion is the explanation for nondepiction, and depiction entails nonocclusion.

An important result of this section will be a theorem that specifies the conditions for completely determining occlusion – we can uniquely determine the extension of the relation *obscures* if and only if there are cues for occlusion, namely, depicted points in occluding edges.

7.2.1 Structures for Occlusion

We will not be expanding our language for $T_{occ}$; rather, we will introduce two defined relations to enhance the readability of the axioms and theorems in this section:

**Definition 7.7** A cue for occlusion at a position is a pixel $q$ that depicts a point in an occluding edge which is in the same convex component of the surfaces as the position $t$.

$$(\forall q, s_1, s_2, t)\text{cue}(q, s_1, s_2, t) \triangleq$$

$$(\exists e, p, t_2)\text{obscures}(s_1, s_2, t) \land \text{occluding}(e, s_2) \land \text{part}(e, s_1) \land \text{part}(p, e) \land \text{position}(p, t_2) \land \Delta(q, p)$$

$$\land \text{interior}(s_1, t_2) \land \text{interior}(s_2, t_2) \land \text{noncrossing}(s_1, t, t_2) \land \text{noncrossing}(s_2, t, t_2)$$

We will later show that the existence of such a cue for occlusion is necessary for disambiguating the extension of *obscures* for depicted scene objects.

**Definition 7.8** A point is a lacuna iff it is neither depicted nor obscured by another surface. 1

$$(\forall p)\text{lacuna}(p) \triangleq (\exists s, t)\text{part}(p, s) \land \text{position}(p, t) \land (\exists q)\Delta(q, p) \land (\exists s')\text{obscures}(s', s, t)$$

The class of intended structures for occlusion and depiction will be represented by $\mathcal{W}_1^{occ}$. The structures in this class will formalize the following intuitions:

**Intuition 25** Occluded points cannot be depicted. Conversely, a point is occluded if the surface of which it is a part, is obscured by another surface at its position.

Since points contain no other scene objects, points cannot be partially depicted. However, edges and surfaces do contain other scene objects, so we have:

**Intuition 26** Edges and surfaces can be partially occluded i.e. there may exist scene objects which are part of the edge or surface which are obscured by another surface, yet the edge or the surface is still depicted.

---

1 The name derives from the Latin word meaning gap, blank space, or missing part. The intuition is that if a point is in the extension of *lacuna*, then there will be a gap in any line which depicts the edge containing the point $p$. 
The definition of \textit{obscures} by itself allows ambiguity in the extension of \textit{obscures}; that is, we are unable to determine whether one surface \textit{obscures} another at an arbitrary point. In fact, in the class of models of $T_{\text{obscures}}$ (i.e. $M_{\text{obscures}}$), there exists a structure for each possible ordering of surfaces at a position. However, if the occluding scene objects are depicted, we can constrain the possible orderings over the set of obscuring surfaces:

\textbf{Intuition 27} We can use cues for occlusion to determine the minimal surfaces in the obscures and ontop orderings over surfaces.

Finally, there is one other relationship between depiction and occlusion which we will need to axiomatize:

\textbf{Intuition 28} If all points in an edge are depicted, then the line depicting the edge will intersect other lines depicting edges in the same surface. If there is a depicted occluding edge for undepicted points in the edge, then the line depicting the edge will intersect other lines depicting occluding edges.

To formally capture this intuition, we will first need to define several preliminary classes of structures.

\textbf{Definition 7.9} Let $M$ be a structure in $\mathcal{L}_{\text{occ}}$. The intersect graph $G = (N, A)$ in $M$ is the substructure of the partial geometry $(I, \in)$ in $M$ consisting of lines such that $(l_1, l_2) \in A$ iff there exists a pixel $q$ such that $(q, l_1), (q, l_2) \in \in$

The intersect graph for the image in Figure 7.7 can be found in Figure 7.8.

\textbf{Definition 7.10} Let $M$ be a structure in $\mathcal{L}_{\text{occ}}$.

The occluding graph $G = (N, A)$ is a directed graph consisting of edges such that $(e_1, e_2) \in A$ iff there exists a surface $s$ such that $(e_1, s) \in \text{occluding}, (e_2, s) \in \text{part}$
Figure 7.8: Intersect graph for the image in Figure 7.7.
Figure 7.9: Occluding and abutting graphs corresponding to the image in Figure 7.7.

and there exist points $p_1, p_2$ and a position $t$ such that

$\langle p_1, e_1 \rangle, \langle p_2, e_2 \rangle \in \text{part}$

$\langle p_1, t \rangle, \langle p_2, t \rangle \in \text{position}$

Definition 7.11 Let $\mathcal{M}$ be a structure in $\mathcal{L}_{\text{occ}}$.

The abutting graph $G = (N, A)$ is a graph consisting of edges such that $(e_1, e_2) \in A$ iff there is a position $t$ such that

$\langle e_1, e_2, t \rangle \in \text{abuts}$

Example: Consider the image $I$ in Figure 7.7. There exists a structure $\mathcal{M} \in \mathcal{L}_{\text{obscures}}$ such that $\mathcal{M} \models I$ and

$\langle l_1, e_1 \rangle, \langle l_2, e_2 \rangle, \langle l_3, e_3 \rangle, \langle l_5, e_5 \rangle, \langle l_7, e_7 \rangle, \langle l_9, e_9 \rangle, \langle l_{11}, e_{11} \rangle \in \Delta$

$\langle l_1, e_1 \rangle, \langle l_2, e_2 \rangle, \langle l_3, e_3 \rangle, \langle l_5, e_5 \rangle, \langle l_7, e_7 \rangle, \langle l_9, e_9 \rangle, \langle l_{11}, e_{11} \rangle \in \Delta$

$\langle s_1, s_2 \rangle, \langle s_1, s_3 \rangle, \langle s_2, s_3 \rangle \in \text{ontop}$

The occluding graph in this structure is shown in Figure 7.9(a), and the abutting graph is shown in Figure 7.9(b).

We can also see how this is related to Intuition 28, in which there are no lacuna points. Lines $l_1$ and $l_5$ intersect, and $l_1$ depicts an occluding edge. Lines $l_1$ and $l_{16}$ intersect, and they depict edges that are part of the same surface. Lines $l_{10}$ and $l_9$ intersect, and they depict abutting edges. $\square$

Following these intuitions, the class of structures corresponding to models of $T_{\text{occ}}$ will need to characterize the relationship between $\text{obscures}$ and the extension of $\Delta$ for each of the classes of scene objects - points, edges, and surfaces.

Definition 7.12 Let $\mathcal{M}_{\text{occ}}$ be a set of structures in $\mathcal{L}_{\text{occ}}$ with the following properties:

For all structures $\mathcal{M} \in \mathcal{M}_{\text{occ}}$,
1. $M$ is an extension of a structure in $M_{obscures}$.

2. $M$ is an extension of a structure in $M_{region}$.

3. For every point $p \in M$ and surface $s \in M$, such that $(p, s) \in \text{part}$, if $(p, t) \in \text{position}$, then $p \in C_{nondep}$ if $s$ is not a minimal surface in the obscures ordering at the position $t$.

4. Given an edge $e \in M$ and surface $s \in M$, such that $(e, s) \in \text{part}$, if $s$ is non-minimal in the obscures ordering at the position of each point in $e$, then $e \in C_{nondep}$.

5. Given a surface $s \in M$, if at every position interior to $s$, it is a non-minimal surface in the obscures ordering at each position, then $s \in C_{nondep}$.

6. If a point $p \in M$ is depicted, then the surface $s$ containing $p$ is minimal in the obscures ordering at every position which is in the same convex component of $s$ as $p$.

7. If an edge $e \in M$ is depicted by a line $l$, then the surface $s$ containing $e$ is minimal in the obscures ordering at the position of each pixel in $l$ that depicts a point in $e$.

8. If a surface $s \in M$ is depicted by a region $r$, then $s$ is minimal in the obscures ordering at each position which is both interior to $s$ and included in $r$.

9. If there are no lacunate points in an edge $e$, then there exists a line $l$ in the intersect graph which depicts $e$, and the edge $e$ is either an element of a contour graph for some surface or it is an element of the occluding graph in $M$, or it is an element of the abutting graph in $M$.

In the definition of this class of structures, condition (3) formalizes Intuition 25, while conditions (4) and (5) formalize Intuition 26. In each of these conditions, we are specifying the relationship between the extension of $obscures$ and elements in $C_{nondep}$ (the undepicted scene objects in a structure $M$). Conditions (6) - (8) formalize Intuition 27, and capture the relationship between the extension of $obscures$ and elements in $C_{dep}$ (the depicted scene objects in a structure $M$). If a scene object is depicted, then there cannot be a surface obscuring all of the scene objects that it contains. Thus, if an edge is depicted, then the surface containing the edge can only be partially occluded. If a region depicts a surface, then the surface cannot be obscured at all positions included in the region.

**Theorem 7.4 Existence Theorem for $M_{occ}$**

The class of structures $M_{occ}$ exists and is nonempty.

**Proof:** Structures in $M_{occ}$ are the union of the following substructures:

1. $obscures$ ordering
2. $C_{nondep}$ (undepicted scene objects)
3. $C_{dep}$ (depicted scene objects)
4. intersect graphs
5. occluding graphs
6. abutting graphs

In earlier chapters, we have shown that the first three classes of structures exist, so that their union exists as well. Conditions (3)-(8) in the definition of $M_{occ}$ place constraints on how these structures can be combined. It is easy to see that for any scene element of $C_{nondep}$, we can construct an $obscures$ ordering in which the surface containing the scene element is non-minimal, and that for any scene element of $C_{dep}$, we can construct an $obscures$ ordering in which the surface containing the scene element is minimal.
If a point is depicted, then it cannot be obscured by a surface. This is equivalent to saying that if a point is obscured by a surface, then it cannot be depicted.

\[(\forall p, q, s, t) \ point(p) \land part(p, s) \land position(p, t) \land \Delta(q, p) \supset (\exists s') \ obscures(s', s, t)\]  

(7.15)

If an edge is depicted, then there exists a point in the edge which is not obscured by a surface. This is equivalent to saying that if all points in an edge are obscured, then the edge cannot be depicted.

\[(\forall l, e, s) \ \Delta(l, e) \land part(e, s) \supset (\exists p, t) \ part(p, e) \land position(p, t) \land (\exists s') \ obscures(s', s, t)\]  

(7.16)

A region depicts a surface iff there exists a position interior to the surface and included in the region which is not obscured by any other surface.

\[(\forall s, r) \ surface(s) \supset \Delta(r, s) \equiv (\exists t) \ interior(s, t) \land inclusion(r, t) \land (\exists s') \ obscures(s', s, t)\]  

(7.17)

If all points in an edge are depicted, then the line depicting the edge will intersect other lines depicting edges in the same surface. If there is a depicted occluding edge for undepicted points in the edge, then the line depicting the edge will intersect other lines depicting occluding edges.

\[(\forall e, s, l) \ edge(e) \land part(e, s) \land \Delta(l, e) \supset [((\exists p) \ point(p) \land part(p, e) \land lacunate(p))\]

\[\lor (\exists l_1, l_2, j_1, j_2, e_1, e_2) \ \Delta(l_1, e_1) \land \Delta(l_2, e_2) \land intersect(l_1, j_1) \land intersect(l_2, j_2) \land (part(e_1, s) \lor occluding(e_1, s) \lor (\exists t) \ abuts(e, e_1, t))\]

\[\land (part(e_2, s) \lor occluding(e_2, s) \lor (\exists t) \ abuts(e, e_2, t))]\]  

(7.18)

---

**7.2.2 Occlusion Axioms**

The axioms in the CardWorld module \( T_{occ} \) are found in Figure 7.10. Axioms 7.15 - 7.17 correspond to conditions in the definition of \( M_{i, obs} \). The existence of occluding and abutting graphs follows from the Satisfiability Theorem for \( M_{i, obs} \), which characterizes the extension of occluding and abuts. Condition (9) in the definition of \( M_{i, occ} \) requires that the union of these three structures exist if there are no lacunate points in an edge. Since each class of structures exists, their union also exists. \( \square \)

---

The existence of intersect graphs follows from the existence of the partial geometry \( I \). The existence of occluding and abutting graphs follows from the Satisfiability Theorem for \( M_{i, obs} \), which characterizes the extension of occluding and abuts. Condition (9) in the definition of \( M_{i, occ} \) requires that the union of these three structures exist if there are no lacunate points in an edge. Since each class of structures exists, their union also exists. \( \square \)

---

The axioms in the CardWorld module \( T_{occ} \) are found in Figure 7.10. Axioms 7.15 - 7.17 correspond to conditions in the definition of \( M_{i, obs} \). In particular, axiom 7.15 captures Intuition 25 - a depicted point cannot be occluded. Axiom 7.16 captures the intuitions about partially occluded edges, while axiom 7.17 captures the intuitions about partially occluded surfaces. This latter axiom is particularly interesting,
since a surface may be depicted even if all of the scene objects that it contains are occluded: all that is required for a surface to be depicted is that there exist a position included in the region at which the surface is not obscured.

Axiom 7.18 of $T_{occ}$ is can be considered to be a depiction axiom for the image relation \textit{intersect}. If all points in an edge are depicted, then the line depicting the edge will intersect other lines depicting edges in the same surface. If there is a depicted occluding edge for undepicted points in the edge, then the line depicting the edge will intersect other lines depicting occluding edges.

\subsection*{7.2.3 Satisfiability of $T_{occ}$}

In this section, we show that the axioms of $T_{occ}$ are consistent, and in particular, that the structures in $\mathcal{M}^i_{occ}$ satisfy the axioms of $T_{occ}$.

\textbf{Theorem 7.5} \textit{Any structure in $\mathcal{M}^i_{occ}$ is a model of $T_{kernel} \cup T_{fg} \cup T_{\Delta} \cup T_{\text{interior}} \cup T_{\Delta}^\text{interior} \cup T_{\text{obscures}} \cup T_{occ} \cup I$, for some planar image $I$.}

\textbf{Proof:} Let $\mathcal{M}$ be a structure in $\mathcal{M}^i_{occ}$.

\[ \mathcal{M} \models (\forall p, q, s, t) \text{point}(p) \land \text{part}(p, s) \land \text{position}(p, t) \land \Delta(q, p) \supset (\exists s') \text{obscures}(s', s, t) \]

iff for any variable assignment $\sigma$, and for all elements $\sigma(p), \sigma(s), \sigma(q), \sigma(t) \in M$, if

\[ \langle \sigma(p) \rangle \in \text{point}, \langle \sigma(p), \sigma(s) \rangle \in \text{part}, \langle \sigma(q), \sigma(p) \rangle \in \Delta, \langle \sigma(p), \sigma(t) \rangle \in \text{position} \]

then there does not exist a surface $s_2$ such that

\[ \langle s_2, \sigma(s), \sigma(t) \rangle \in \text{obscures} \]

so that $\sigma(s)$ is a minimal surface in the \textit{obscures} ordering at the position $\sigma(t)$, which follows from condition (3) in the definition of $\mathcal{M}^i_{occ}$, since by the premise we have $\sigma(p) \notin C_{\text{nondep}}$.

Thus, axiom 7.15 of $T_{occ}$ is satisfied in any structure in $\mathcal{M}^i_{occ}$.

\[ \mathcal{M} \models (\forall l, e, s) \Delta(l, e) \land \text{part}(e, s) \supset (\exists p, t) \text{part}(p, e) \land \text{position}(p, t) \land (\exists s') \text{obscures}(s', s, t) \]

iff for any variable assignment $\sigma$, and for all elements $\sigma(l), \sigma(e), \sigma(s) \in M$, if

\[ \langle \sigma(l), \sigma(e) \rangle \in \Delta, \langle \sigma(e), \sigma(s) \rangle \in \text{part} \]

then there does not exist an element $s_2 \in M$ such that

\[ \langle s_2, \sigma(s), \sigma(t) \rangle \in \text{obscures} \]

so that $\sigma(s)$ is a minimal surface in the \textit{obscures} ordering at each position $\sigma(t)$, which follows from condition (4) in the definition of $\mathcal{M}^i_{occ}$, since by the premise we have $\sigma(e) \notin C_{\text{nondep}}$.

Thus, axiom 7.16 of $T_{occ}$ is satisfied.

\[ \mathcal{M} \models (\forall s, r) \text{surface}(s) \supset \Delta(r, s) \equiv ((\exists t) \text{interior}(s, t) \land \text{inclusion}(r, t) \land (\exists s') \text{obscures}(s', s, t)) \]

iff for any variable assignment $\sigma$, and for all elements $\sigma(s), \sigma(r) \in M$, we have

\[ \langle \sigma(s) \rangle \in \text{surface}, \langle \sigma(r), \sigma(s) \rangle \in \Delta \]
iff there exists a position $t \in M$ such that

$$(\sigma(s), t) \in \text{interior}, (\sigma(t), t) \in \text{inclusion}$$

then there does not exist a surface $s_2 \in M$ such that

$$(s_2, \sigma(s), t) \in \text{obsures}$$

so that $\sigma(s)$ is a minimal surface in the obsures ordering at each position $\sigma(t)$, which follows from condition (5) in the definition of $M_i^{\text{occ}}$, since by the premise we have $\sigma(e) \notin C_{\text{nondep}}$.

Thus, axiom 7.17 of $T_{\text{occ}}$ is satisfied.

$$\mathcal{M}, \sigma \models (\forall e, s, l) \ (\text{edge}(e) \land \text{part}(e, s) \land \Delta(l, e) \supset [(\exists p) \ (\text{point}(p) \land \text{part}(p, e) \land \text{lacunate}(p))]

\lor (\exists l_1, l_2, j_1, j_2, e_1, e_2) \ (\Delta(l_1, e_1) \land \Delta(l_2, e_2) \land \text{intersect}(l_1, l_1, j_1) \land \text{intersect}(l_1, l_2, j_2)

\land (\text{part}(e_1, s) \lor \text{occluding}(e_1, s) \lor (\exists t) \ \text{abuts}(e, e_1, t))

\land (\text{part}(e_2, s) \lor \text{occluding}(e_2, s) \lor (\exists t) \ \text{abuts}(e, e_2, t)))$$

iff for any variable assignment $\sigma$, and for every edge $\sigma(e)$ depicted by a line $\sigma(l)$, either there exists a lacuna point, or there exist edges $\sigma(e_1), \sigma(e_2)$, lines $\sigma(l_1), \sigma(l_2)$ and pixels $\sigma(j_1), \sigma(j_2)$ such that

$$(\exists l_1, l_2, j_1, j_2, e_1, e_2) \ (\text{part}(e_1, s) \lor \text{occluding}(e_1, s) \lor (\exists t) \ \text{abuts}(e, e_1, t))$$

and

$$(\exists l_1, l_2, j_1, j_2, e_1, e_2) \ (\text{part}(e_2, s) \lor \text{occluding}(e_2, s) \lor (\exists t) \ \text{abuts}(e, e_2, t))$$

The first sentence is satisfied iff $\sigma(l)$ is an element of the intersect graph.

The second sentence is satisfied iff

$$(e_1, s) \in \text{part} \lor (e_1, s) \in \text{occluding} \lor (e, e_1, t) \in \text{abuts}$$

which follows from property (9) in the definition of $M_i^{\text{occ}}$.

The third sentence is satisfied iff

$$(e_2, s) \in \text{part} \lor (e_2, s) \in \text{occluding} \lor (e, e_2, t) \in \text{abuts}$$

which also follows from property (9) in the definition of $M_i^{\text{occ}}$.

Thus, axiom 7.18 of $T_{\text{occ}}$ is satisfied. $\square$

### 7.2.4 Cues for Occlusion

In this section, we specify the conditions necessary to determine the complete extension of obsures for occluded scene objects when we are given an image. It will be a key result needed to prove the Axiomatizability Theorem for occlusion, since it corresponds to conditions (6) - (8) in the definition of $M_i^{\text{occ}}$, which in turn correspond to Intuition 27.

The problem with which we are faced is the need to disambiguate obsures when we are given the positions of all points in two surfaces. Using only $T_{\text{obsures}}$, scenes are ambiguous with respect to the extension of obsures. Given two surfaces, all we can say is that one obscures the other: we cannot even determine which surface is minimal in the obsures ordering. In fact, the definition of $M_i^{\text{obsures}}$
Figure 7.11: Cues for occlusion.
shows that for any surface, there exists a structure in $M_1^{\text{obscures}}$ in which that surface is minimal in the \textit{obscures} ordering at some position.

There are other cases that are consistent with the occlusion axioms but which also are ambiguous with respect to \textit{obscures}. If there are lacunate errors in the edge detection as with the image in Figure 7.11(a), in which we have a silhouette of a set of surfaces, we may still be unable to decide \textit{obscures}. The same scene with hidden edges displayed is in Figure 7.11(b). Similarly, in Figure 7.11(c) and (e) we do not know the extension of \textit{obscures} for the surfaces depicted by $r_2$ and $r_3$, since the place of occlusion is itself occluded. These scenes with hidden edges displayed are in Figures 7.11(d) and (f). However, in these cases it doesn't matter which surface is obscuring and which is obscured; there is no evidence to distinguish one assignment from another, and any assignment is consistent. We therefore want to distinguish the cases where the assignment of \textit{obscures} is completely determined.

Using the notion of convex components, axiom 7.8 of $T_{\text{obscures}}$ can be thought of as requiring any two positions in the intersection of the convex components of two surfaces have the same \textit{obscures} ordering. The parts of the scene where there is ambiguity over \textit{obscures} are those positions that are \textit{interior} to multiple surfaces. The following two theorems, which are the characterization theorems for \textit{obscures}, define the conditions under which we can remove this ambiguity.

We previously presented the definition of occluding edges and proved that occluding edges exist. We now consider models in which the occluding edge is depicted. The definition of $\text{cue}(q, s_1, s_2, t)$ is used to demonstrate that depicted occluding edges provide information about the \textit{obscures} ordering for surfaces in the scene. The motivation behind this definition is to define what constitutes evidence for occlusion. A pixel $q$ is a cue for the occlusion of $s'$ by $s$ at $t$ iff $s$ obscures $s'$ at some position $t$ and the occluding edge for this position contains a point that is depicted by $q$. This pixel provides evidence for occlusion since by axiom 7.15 a depicted scene object cannot be occluded; if there is any ambiguity with respect to \textit{obscures} between two surfaces, then the depicted surface is the unobscured one. Notice that in the examples considered above, ambiguity existed in images where there were no cues for occlusion. The following theorems make these intuitions precise.

**Theorem 7.6** For any planar image $I$, if

$$T_{\text{scene}} \cup T_{\text{kernel}} \cup T_{fg} \cup T_{\Delta} \cup T_{\text{interior}} \cup T_{\text{interior}}^{\Delta} \cup T_{\text{obscures}} \cup T_{\text{occ}} \cup I \models$$

$$\exists s, s', t) \text{interior}(s, t) \land \text{interior}(s', t) \land \neg(\exists q) \text{cue}(q, s, s', t) \lor \text{cue}(q, s', s, t)$$

then there exist models $M_1, M_2$ of $T_{\text{scene}} \cup T_{\text{kernel}} \cup T_{fg} \cup T_{\Delta} \cup T_{\text{interior}} \cup T_{\text{interior}}^{\Delta} \cup T_{\text{obscures}} \cup T_{\text{occ}} \cup I$ such that for any variable assignment $\sigma$ we have

$$M_1, \sigma \models \text{obscures}(s, s', t)$$

$$M_2, \sigma \models \text{obscures}(s', s, t)$$

In other words, if there are no cues for occlusion for two surfaces in the image, then any ordering is consistent.
Proof: Suppose

\[ T_{\text{scene}} \cup T_{\text{kernel}} \cup T_{\text{fg}} \cup T_{\Delta} \cup T_{\text{interior}} \cup T_{\text{interior}}^\Delta \cup T_{\text{obscures}} \cup T_{\text{occ}} \cup I \models \]

\[ (\exists s, s', t) \text{interior}(s, t) \land \text{interior}(s', t) \land (\neg (\exists q) \text{cue}(q, s, s', t) \lor \text{cue}(q, s', s, t)) \land \text{obscures}(s, s', t) \]

The only axiom in \( T_{\text{occ}} \) that can be violated is axiom 7.15, so that we have

\[ T_{\text{kernel}} \cup T_{\text{fg}} \cup T_{\Delta} \cup T_{\text{interior}} \cup T_{\text{interior}}^\Delta \cup T_{\text{obscures}} \cup T_{\text{occ}} \cup I \models \]

\[ (\exists s', t, p, q) \text{point}(p) \land \text{part}(p, s) \land \text{position}(p, t) \land \Delta(q, p) \land \text{obscures}(s', s, t) \]

but in this case, we would have a cue for the occlusion of \( s \) by \( s' \), which contradicts the hypothesis that there are no cues for the occlusion of either surface. \( \square \)

**Theorem 7.7** Let \( \mathcal{M}, \mathcal{M}' \) be models of \( T_{\text{scene}} \cup T_{\text{kernel}} \cup T_{\text{fg}} \cup T_{\Delta} \cup T_{\text{interior}} \cup T_{\text{interior}}^\Delta \cup T_{\text{obscures}} \cup T_{\text{occ}} \cup I \) for some planar image \( I \). Suppose for some variable assignment \( \sigma \),

\[ \mathcal{M}, \sigma \models \text{interior}(s, t) \land \text{interior}(s', t) \supset (\exists q) \text{cue}(q, s, s', t) \lor \text{cue}(q, s', s, t) \]

iff

\[ \mathcal{M}', \sigma \models \text{interior}(s, t) \land \text{interior}(s', t) \supset (\exists q) \text{cue}(q, s, s', t) \lor \text{cue}(q, s', s, t) \]

and

\[ \mathcal{M}, \sigma \models \text{noncrossing}(s, t, t_2) \land \text{noncrossing}(s', t, t_2) \]

iff

\[ \mathcal{M}', \sigma \models \text{noncrossing}(s, t, t_2) \land \text{noncrossing}(s', t, t_2) \]

Then

\[ \mathcal{M}, \sigma \models \text{obscures}(s, s', t_2) \]

iff

\[ \mathcal{M}', \sigma \models \text{obscures}(s, s', t_2) \]

In other words, if there are cues for occlusion in the image, then all models must agree on the \text{obscures} ordering at positions which are in the same convex component as the position of the cue.

**Proof:** Suppose \( \mathcal{M} \) and \( \mathcal{M}' \) are models of \( T_{\text{scene}} \cup T_{\text{kernel}} \cup T_{\text{fg}} \cup T_{\Delta} \cup T_{\text{interior}} \cup T_{\text{interior}}^\Delta \cup T_{\text{obscures}} \cup T_{\text{occ}} \cup I \) with the same domain, and \( \sigma \) is a variable assignment such that

\[ \mathcal{M}, \sigma \models \text{interior}(s_1, t_1) \land \text{interior}(s_2, t_1) \land (\text{cue}(q_1, s_1, s_2, t_1) \lor \text{cue}(q_1, s_2, s_1, t_1)) \land \text{noncrossing}(s_1, t, t_2) \land \text{noncrossing}(s_2, t, t_2) \]

iff

\[ \mathcal{M}', \sigma \models \text{interior}(s_1, t_1) \land \text{interior}(s_2, t_1) \land (\text{cue}(q_1, s_1, s_2, t_1) \lor \text{cue}(q_1, s_2, s_1, t_1)) \land \text{noncrossing}(s_1, t, t_2) \land \text{noncrossing}(s_2, t, t_2) \]

Since the line \( q_1 \) in the cue for occlusion depicts a point in an occluding edge, by axioms 7.1, 7.2, and 7.4 of \( T_{\text{obscures}} \) if

\[ \mathcal{M}, \sigma \models \text{interior}(s_1, t_1) \land \text{interior}(s_2, t_1) \]

then

\[ \mathcal{M}, \sigma \models \text{obscures}(s_1, s_2, t_1) \equiv \neg \text{obscures}(s_2, s_1, t_1) \]
and if
\[ M', \sigma \models \text{interior}(s_1, t_1) \land \text{interior}(s_2, t_1) \]
then
\[ M', \sigma \models \text{obscures}(s_1, s_2, t_1) \equiv \neg \text{obscures}(s_2, s_1, t_1) \]

By the Occluding Edge Existence Lemma, we have
\[ M, \sigma \models (\exists e, p, l, q, t) \text{part}(e, s) \land \text{occluding}(e, s_2) \land \Delta(l, e) \land \text{part}(p, e) \land \text{position}(p, t) \land \Delta(q, p) \]
niff
\[ M', \sigma \models (\exists e, p, l, q, t) \text{part}(e, s) \land \text{occluding}(e, s_2) \land \Delta(l, e) \land \text{part}(p, e) \land \text{position}(p, t) \land \Delta(q, p) \]
or
\[ M, \sigma \models \neg (\exists e, p, l, q, t) \text{part}(e, s) \land \text{occluding}(e, s_2) \land \Delta(l, e) \land \text{part}(p, e) \land \text{position}(p, t) \land \Delta(q, p) \]
niff
\[ M', \sigma \models \neg (\exists e, p, l, q, t) \text{part}(e, s) \land \text{occluding}(e, s_2) \land \Delta(l, e) \land \text{part}(p, e) \land \text{position}(p, t) \land \Delta(q, p) \]

If we put these together we get
\[ M, \sigma \models \text{obscures}(s_1, s_2, t) \]
niff
\[ M', \sigma \models \text{obscures}(s_1, s_2, t) \]
or
\[ M, \sigma \models \text{obscures}(s_2, s_1, t) \]
niff
\[ M', \sigma \models \text{obscures}(s_2, s_1, t) \]
and by axiom 7.8, we have
\[ M, \sigma \models \text{obscures}(s_1, s_2, t_2) \]
niff
\[ M', \sigma \models \text{obscures}(s_1, s_2, t_2) \]
or
\[ M, \sigma \models \text{obscures}(s_2, s_1, t_2) \]
niff
\[ M', \sigma \models \text{obscures}(s_2, s_1, t_2) \]

Thus, if we have cues for occlusion for every obscured surface, we can uniquely determine the minimal surface in the \text{obscures} ordering. If we do not have cues for occlusion for some obscured surface, then there exists a model for every local surface ordering (as we had with \( T_{\text{obscures}} \) alone).

It is important to note that these theorems are independent of any assumptions about errors in edge detection, as the following example illustrates.
Figure 7.12: Relationship between cues for occlusion and errors in edge detection.
Example: Let $I$ be the image in Figure 7.12(a). The scene with hidden edges displayed is illustrated in Figure 7.12(b). There is a model $\mathcal{M}$ of $T_{\text{scene}} \cup T_{\text{kernel}} \cup T_{fg} \cup T_{\Delta} \cup T_{\text{interior}} \cup T_{\text{interior}}^\Delta \cup T_{\text{obscures}} \cup I$ in which there are errors in edge detection, yet we can completely decide $\text{obscures}$ for the surfaces, since $l_1$ depicts an edge occluding the surface $s_3$ and $l_2$ depicts an edge occluding the surface $s_1$.

Example: Let $I$ be the image in Figure 7.12(c). Consider a model $\mathcal{M}$ of $T_{\text{scene}} \cup T_{\text{kernel}} \cup T_{fg} \cup T_{\Delta} \cup T_{\text{interior}} \cup T_{\text{interior}}^\Delta \cup T_{\text{obscures}} \cup I$ that satisfies

$$(r_1, s_1), (r_2, s_2), (r_3, s_3) \in \Delta$$

The scene with hidden edges displayed is illustrated in figure 7.12(d). In this model, we cannot determine whether $s_2$ is on top of $s_3$ or whether $s_3$ is on top of $s_2$ since no occluding edges are depicted.

Further, there is a model that satisfies

$$(r_1, s_1), (r_2, s_2), (r_3, s_3), (r_3, s_2) \in \Delta$$

$$(s_2, s_3) \in \text{ontop}$$

(in which case there are errors in edge detection). The image in Figure 7.12(e) is the same scene, but in this case there are no errors in edge detection and the occluding edge is depicted. □

### 7.2.5 Axiomatizability Theorem for Occlusion

In this section, we show that $\mathcal{M}_i^{\text{occ}}$ characterizes the models of $T_{\text{occ}}$.

**Theorem 7.8** Let $\mathcal{M}$ be a structure in $L_{\text{occ}}$ with the following properties:

- $\mathcal{M}$ contains a substructure isomorphic to a structure in $\mathcal{M}_i^{\text{Polygon}}$;
- $\mathcal{M}$ contains a substructure isomorphic to a structure in $\mathcal{M}_i^{\text{Angle}}$;
- $\mathcal{M}$ contains a substructure isomorphic to a structure in $\mathcal{M}_i^{\text{Plane}}$;
- $\mathcal{M}$ contains a substructure isomorphic to a structure in $\mathcal{M}_i^{\text{Planar}}$ (i.e. any image $I$ are planar);
- $\mathcal{M}$ is a model of $T_{\text{kernel}} \cup T_{fg} \cup T_{\Delta} \cup T_{\text{interior}} \cup T_{\text{interior}}^\Delta \cup T_{\text{obscures}} \cup T_{\text{occ}} \cup I$ for some planar image $I$.

Then $\mathcal{M}$ is isomorphic to a structure in $\mathcal{M}_i^{\text{occ}}$.

**Proof:** Suppose that $\mathcal{M}$ is a model of $T_{\text{kernel}} \cup T_{fg} \cup T_{\Delta} \cup T_{\text{interior}} \cup T_{\text{interior}}^\Delta \cup T_{\text{obscures}} \cup T_{\text{occ}} \cup I$ for some planar image $I$, and that it satisfies the properties in the hypothesis of the theorem.

The first two conditions in the definition of $\mathcal{M}_i^{\text{occ}}$ follow from the hypothesis of the theorem. Consider axiom 7.15 of $T_{\text{occ}}$:

\[ \mathcal{M} \models (\forall p, q, s, t) \text{point}(p) \land \text{part}(p, s) \land \text{position}(p, t) \land \Delta(q, p) \supset \neg(\exists s') \text{obscures}(s', s, t) \]
This is equivalent to
\[ M \models (\forall p, s_1, s_2, t) \text{ point}(p) \land \text{part}(p, s) \land \text{position}(p, t) \land \text{obscures}(s_2, s_1, t) \supset (\exists q) \Delta(q, p) \]
For any variable assignment \( \sigma \), and for any elements \( \sigma(p), \sigma(s_1), \sigma(s_2), \sigma(t) \in M \), if
\[ \langle \sigma(p) \rangle \in \text{point}, \langle \sigma(p), \sigma(t) \rangle \in \text{position}, \langle \sigma(s_2), \sigma(s_1), \sigma(t) \rangle \in \text{obscures} \]
then there does not exist \( q \in M \) such that \( \langle q, \sigma(p) \rangle \in \Delta \).
Since \( M \) is a model of \( T_{obscures} \), the antecedent is true iff \( \sigma(s_1) \) is not a minimal element of the \( \text{obscures} \) ordering at \( \sigma(t) \). Since \( M \) is a model of \( T_{kernel} \), the consequent is true iff \( \sigma(p) \notin C_{nondep} \).
Thus, condition (3) in the definition of \( M_i^{occ} \) is satisfied by \( M \).
Consider axiom 7.16 of \( T_{occ} \):
\[ M \models (\forall l, e, s) \Delta(l, e) \land \text{part}(e, s) \supset (\exists p, t) \text{part}(p, e) \land \text{position}(p, t) \land (\exists s') \text{obscures}(s', s, t) \]
This is equivalent to
\[ M \models (\forall e, s_1) \text{part}(e, s_1) \land ((\forall p, t) \text{part}(p, e, s_1) \land \text{part}(p, e) \land \text{position}(p, t) \supset (\exists s_2) \text{obscures}(s_2, s_1, t)) \]
\[ \supset (\exists l) \Delta(l, e) \]
so that for any variable assignment \( \sigma \), and for any elements \( \sigma(e), \sigma(s_1), \sigma(p), \sigma(t) \in M \), if
\[ \langle \sigma(e), \sigma(s_1) \rangle, \langle \sigma(p), \sigma(t) \rangle \in \text{part}, \langle \sigma(p), \sigma(t) \rangle \in \text{position} \]
implies the existence of a surface \( s_2 \) such that
\[ \langle s_2, \sigma(s_1), \sigma(t) \rangle \in \text{obscures} \]
then there does not exist a line \( l \in M \) such that
\[ \langle l, \sigma(e) \rangle \in \Delta \]
Since \( M \) is a model of \( T_{obscures} \), the antecedent is true iff at the position of each point in \( \sigma(e) \) the surface \( \sigma(s_1) \) is not minimal in the \( \text{obscures} \) ordering at that position. Since \( M \) is a model of \( T_{kernel} \), the consequent is true iff \( \sigma(e) \notin C_{nondep} \). Thus, condition (4) in the definition of \( M_i^{occ} \) is satisfied by \( M \).
Consider axiom 7.17 of \( T_{occ} \):
\[ M \models (\forall s, r) \text{surface}(s) \supset \Delta(r, s) \equiv ((\exists t) \text{interior}(s, t) \land \text{inclusion}(r, t) \land (\exists s') \text{obscures}(s', s, t)) \]
This is equivalent to
\[ M \models (\forall s) [\text{surface}(s) \land ((\forall r, t) \text{interior}(s, t) \land \text{inclusion}(r, t) \supset (\exists s') \text{obscures}(s', s, t))] \supset (\exists s') \text{obscures}(s', s, t) \]
and
\[ M \models (\forall s, r, t) [\text{surface}(s) \land \Delta(r, s) \land \text{interior}(s, t) \land \text{inclusion}(r, t) \supset (\exists s') \text{obscures}(s', s, t)] \]
By the first sentence, for any variable assignment \( \sigma \), if for all regions \( \sigma(r) \in M \) and all positions \( \sigma(t) \in M \) which are interior to \( \sigma(s) \in M \) and included in \( \sigma(r) \), there exists an obscuring surface at that position, then
\[ \langle \sigma(r), \sigma(s) \rangle \notin \Delta \]
Since this holds for all regions in \( M \), we have \( \sigma(s) \in C_{nondep} \).
Thus, condition (5) in the definition of \( M_i^{occ} \) is satisfied by \( M \).
By the second sentence, if there exists a position \( \sigma(t) \) such that
\[
(\sigma(s), \sigma(t)) \in \text{interior}, (\sigma(r), \sigma(t)) \in \text{inclusion}
\]
then there does not exist a surface \( s_2 \in M \) such that
\[
(s_2, \sigma(s), t) \in \text{obscures}
\]
so that \( \sigma(s) \) is the minimal surface in the obscures ordering at \( \sigma(t) \).
Thus, condition (8) in the definition of \( M_{\text{occ}} \) is satisfied by \( \mathcal{M} \).
By Theorem 7.7, we have
\[
\mathcal{M}, \sigma \vdash (\forall q, p, e, s_1, s_2, t) \text{cue}(q, s_1, s_2, t) \wedge \Delta(q, p) \wedge \text{part}(p, e) \wedge \text{position}(p, t) \supset \text{obscures}(s_1, s_2, t)
\]
By axiom 7.15 of \( T_{\text{occ}} \), we get for any variable assignment \( \sigma \),
\[
\mathcal{M}, \sigma \models \neg(\exists s) \text{obscures}(s, s_1, t)
\]
so that \( \sigma(s) \) is the minimal in the obscures ordering at the position \( \sigma(t) \). By axiom 7.8 of \( T_{\text{obscures}} \), it is minimal at every position in the same convex component of \( \sigma(s) \) as \( \sigma(t) \).
Thus condition (6) in the definition of \( M_{\text{occ}}^1 \) is satisfied by \( \mathcal{M} \).
Suppose
\[
\mathcal{M}, \sigma \models (\exists l) \Delta(l, e)
\]
for some edge \( \sigma(e) \). By \( T_{\text{kernel}} \), we have
\[
\mathcal{M}, \sigma \vdash (\forall l) \Delta(l, e) \supset (\exists q, p, s) \text{in}(q, l) \wedge \Delta(q, p) \wedge \text{part}(e, s)
\]
By Theorem 7.7, we have
\[
\mathcal{M}, \sigma \vdash (\forall q, p, s, t) \text{cue}(q, s, t) \wedge \Delta(q, p) \wedge \text{part}(p, e) \wedge \text{position}(p, t) \supset \text{obscures}(s_1, s_2, t)
\]
By axiom 7.15 of \( T_{\text{occ}} \), we get
\[
\mathcal{M}, \sigma \models \neg(\exists s) \text{obscures}(s_1, t)
\]
so that \( \sigma(s_1) \) is the minimal in the obscures ordering at the position \( \sigma(t) \). By axiom 7.8 of \( T_{\text{obscures}} \), it is minimal at every position in the same convex component of \( \sigma(s) \) as \( \sigma(t) \).
Thus condition (7) in the definition of \( M_{\text{occ}}^1 \) is satisfied by \( \mathcal{M} \).
Suppose, for any variable assignment \( \sigma \),
\[
\mathcal{M}, \sigma \models (\forall e, s, l) \text{edge}(e) \wedge \text{part}(e, s) \wedge \Delta(l, e) \supset [(\exists p) \text{point}(p) \wedge \text{part}(p, e) \wedge \text{lacunate}(p)]
\]
\[
\vee (\exists l_1, l_2, j_1, j_2, e_1, e_2) \Delta(l_1, e_1) \wedge \Delta(l_2, e_2) \wedge \text{intersect}(l_1, l_2, j_1) \wedge \text{intersect}(l_1, l_2, j_2)
\]
\[
\wedge (\text{part}(e_1, s)) \vee \text{occluding}(e_1, s) \vee (\exists t) \text{abuts}(e, e_1, t)
\]
\[
\wedge (\text{part}(e_2, s) \vee \text{occluding}(e_2, s) \vee (\exists t) \text{abuts}(e, e_2, t)))
\]
Then for every edge \( \sigma(e) \) depicted by a line \( \sigma(l) \), either there exists a lacuna point, or there exist edges \( \sigma(e_1), \sigma(e_2) \), lines \( \sigma(l_1), \sigma(l_2) \) and pixels \( \sigma(j_1), \sigma(j_2) \) such that
\[
\mathcal{M}, \sigma \models \Delta(l_1, e_1) \wedge \Delta(l_2, e_2) \wedge \text{intersect}(l_1, l_1, j_1) \wedge \text{intersect}(l_1, l_2, j_2)
\]
and
\[
\mathcal{M}, \sigma \models (\text{part}(e_1, s) \vee \text{occluding}(e_1, s) \vee (\exists t) \text{abuts}(e, e_1, t))
\]
and
\[
\mathcal{M}, \sigma \models (\text{part}(e_2, s) \vee \text{occluding}(e_2, s) \vee (\exists t) \text{abuts}(e, e_2, t)))
\]
The first sentence is satisfied iff $\sigma(l)$ is an element of the intersect graph.

The second sentence is satisfied iff

$\langle e_1, s \rangle \in \text{part} \text{ or } \langle e_1, s \rangle \in \text{occluding} \text{ or } \langle e, e_1, t \rangle \in \text{abuts}$

while the third sentence is satisfied iff

$\langle e_2, s \rangle \in \text{part} \text{ or } \langle e_2, s \rangle \in \text{occluding} \text{ or } \langle e, e_2, t \rangle \in \text{abuts}$

so that $\mathcal{M}$ satisfies property (9) in the definition of $\mathcal{M}_i^{\text{occ}}$.

Therefore, $\mathcal{M}$ satisfies all of the properties in the definition of $\mathcal{M}_i^{\text{occ}}$, and hence $\mathcal{M} \in \mathcal{M}_i^{\text{occ}}$. $\square$

In the next chapter we explore the relationship between occlusion and depiction in more depth.
Chapter 8

Reasoning about Occlusion with No Errors in Edge Detection

In the preceding chapters we have proven characterization theorems for the eight modules of the CardWorld theories. In this chapter we employ a new methodology—we introduce various sentences as assumptions that are independent of the CardWorld axioms. We will be interested in proving properties of these assumptions and the relationships between them. In particular, we will make the assumption that occlusion and nondepiction are equivalent, that is, we make the assumption that every undepicted scene object is occluded by some surface. This is actually an assumption about errors in edge detection, since with such errors there may exist undepicted scene objects which are not occluded. We then show that this occlusion assumption is logically equivalent to a set of independent depiction assumptions. The relationships between the occlusion assumption and the depiction assumptions provide a complete characterization of the assumption about errors in edge detection.

This characterization also provides sufficient conditions for transforming the CardWorld axioms into a logically equivalent set of propositional theories, laying the groundwork for the complexity results in the final chapter of the thesis. The transformation from an occlusion assumption to a set of depiction assumptions is crucial for the computational aspect of this work, since we will primarily be concerned with reasoning problems that determine which regions depict surfaces, and which lines depict edges in those surfaces.

In the final section of this chapter, we will show how the characterization theorems of the preceding chapters can be strengthened by the occlusion and depiction assumptions.

8.1 Errors in Edge Detection: Overview

Before presenting the axioms for occlusion we need to examine the difference between occlusion and nondepiction. A scene object can be nondepicted for two reasons—either there is a surface on top of the

\[ T_{\text{cardworld}} = T_{\text{scene}} \cup T_{\text{kernel}} \cup T_{\text{fg}} \cup T_{\Delta} \cup T_{\text{interior}} \cup T_{\text{interior}} \cup T_{\text{obsures}} \cup T_{\text{occ}}. \]

---

\(1\) For the remainder of the thesis, we will use \( T_{\text{cardworld}} \) to denote the theory \( T_{\text{scene}} \cup T_{\text{kernel}} \cup T_{\text{fg}} \cup T_{\Delta} \cup T_{\text{interior}} \cup T_{\text{interior}} \cup T_{\text{obsures}} \cup T_{\text{occ}}.\)
object that is occluding it, or it is not depicted because of errors in edge detection. Such errors occur when pieces of an edge are not depicted due to poor segmentation (see Figure 8.1(a),(b),(c),(e)). We will call these kinds of errors lacunate errors, since there are gaps in the lines depicting the edges.

Example: Let \( I \) be the image in Figure 8.1(c). There is a model \( \mathcal{M} \) of \( T_{\text{cardworld}} \cup \mathcal{I} \) such that
\[
\{e_1, e_2, e_3, e_4, s_1\} \subset M
\]
and
\[
\langle l_1, e_1 \rangle, \langle l_2, e_1 \rangle, \langle l_3, e_2 \rangle, \langle l_4, e_3 \rangle, \langle l_5, e_4 \rangle \in \Delta
\]
\[
\langle e_1, s_1 \rangle, \langle e_2, s_1 \rangle, \langle e_3, s_1 \rangle, \langle e_4, s_1 \rangle \in \text{part}
\]

In this model, there are "gaps" and missing pieces in the lines depicting the edges of the surface. The scene with all unoccluded scene objects depicted is illustrated in Figure 8.1(f). \( \square \)

Example: Let \( I \) be the image in Figure 8.1(e). There is a model \( \mathcal{M} \) of \( T_{\text{cardworld}} \cup \mathcal{I} \) such that
\[
\{e_1, e_2, e_3, e_4, s_1, s_2, s_3\} \subset M
\]
and
\[
\langle l_1, e_1 \rangle, \langle l_2, e_2 \rangle, \langle l_3, e_3 \rangle, \langle l_4, e_4 \rangle \in \Delta
\]
\[
\langle r_1, s_1 \rangle, \langle r_1, s_2 \rangle, \langle r_1, s_3 \rangle \in \Delta
\]
\[
\langle e_1, s_1 \rangle, \langle e_2, s_1 \rangle, \langle e_3, s_2 \rangle, \langle e_4, s_2 \rangle \in \text{part}
\]
\[
\langle s_1, s_2 \rangle, \langle s_2, s_3 \rangle \in \text{ontop}
\]

In this model, there are "gaps" and missing pieces in the lines depicting the edges of the occluding surfaces. \( \square \)

Thus, in general, \( T_{\text{cardworld}} \cup \mathcal{I} \) is consistent with images in which there exist undepicted scene objects that are not occluded. However, we will now restrict ourselves to the special case of object recognition in which there are no lacunate errors in edge detection.

### 8.1.1 Assumptions about Errors in Edge Detection

To eliminate the models in which there exist undepicted scene objects which are unoccluded, we introduce the following assumption:

**Definition 8.1** The No Lacunate Errors Assumption (NLEA) assumption is the following sentence
\[
(\forall p, s, t) \ point(p) \land \ part(p, s) \land \ position(p, t) \supset [(\exists q) \ Delta(q, p) \equiv \neg(\exists s') \ obsures(s', s, t)]
\]
\[
i.e., \ a \ point \ is \ depicted \ iff \ it \ is \ unobscured. \ Equivalently, \ a \ point \ is \ undepicted \ iff \ it \ is \ obscured.
\]

\(^2\)This term is derived from the word "lacuna" meaning gap, blank space, or missing part.
Figure 8.1: Errors in edge detection.
Alternatively, we can use the *lacuna* relation from $T_{occ}$ to rewrite this as

$$(\forall p) \text{point}(p) \supset \neg \text{lacuna}(p)$$

In the remainder of the thesis, we will restrict ourselves to models of $T_{cardworld}$ which satisfy this assumption; that is, we will consider models of $T_{cardworld} \cup NLEA \cup I$.

In axiom 7.15 of $T_{occ}$, all occluded scene objects are undepicted: with NLEA, all undepicted scene objects are occluded. Also note that there are two ways to violate NLEA — either there is no surface *ontop* of the nondepicted scene object, or there is an occluding surface but it is not depicted.

**Example:** Let $I$ be the image in Figure 8.1(a). There is a model $\mathcal{M}$ of $T_{cardworld} \cup I$ such that

$${e_1, e_2, e_3, s_1, s_2} \subseteq M$$

$$(l_1, e_1), (l_2, e_2), (l_3, e_3), (l_4, e_4) \in \Delta$$

$$(e_1, s_1), (e_2, s_1), (e_3, s_1), (e_4, s_1) \in \text{part}$$

The scene with all undepicted scene objects displayed is illustrated in Figure 8.1(f): in this case there is only one surface.

There is also a model $\mathcal{N}$ of $T_{cardworld} \cup I$ that satisfies:

$$(l_1, e_1), (l_2, e_2), (l_3, e_3), (l_4, e_4) \in \Delta$$

$$(e_2, s_1), (e_3, s_1), (e_4, s_1) \in \text{part}$$

$$(s_1, s_1) \in \text{ontop}$$

The scene with all undepicted scene objects displayed is illustrated in Figure 8.1(g): in this case there are two surfaces, so that the undepicted scene objects in the surface are due to occlusion but the occluding surface is undepicted due to lacunate errors. □

There is another kind of error with edge detection that arises when lines or pixels are present that do not depict any scene object: we will call these spurious errors (see Figure 8.1(d)).

**Example:** Let $I$ be the image in Figure 8.1(d). There is a model $\mathcal{M}$ of $T_{cardworld} \cup I$ such that

$${e_1, e_2, e_3, e_4, s_1} \subseteq M$$

$$(l_1, e_1), (l_3, e_1), (l_4, e_2), (l_6, e_2), (l_8, e_3), (l_{14}, e_3), (l_{13}, e_4) \in \Delta$$

$$(e_1, s_1), (e_2, s_1), (e_3, s_1), (e_4, s_1) \in \text{part}$$

In this model, the lines $l_2, l_{11}, l_5, l_7, l_9, l_{12}$, and $l_{10}$ do not depict any edges. □

We therefore introduce an assumption which eliminates such models, so that every pixel and line depicts some scene object:
Definition 8.2 The No Spurious Errors Assumption (NSEA) is the set of sentences
\[
(\forall q) \text{pixel}(q) \supset (\exists p) \text{point}(p) \land \Delta(q, p) \\
(\forall l) \text{line}(q) \supset (\exists e) \text{edge}(e) \land \Delta(l, e)
\]
i.e., every pixel depicts a point, and every line depicts an edge.

This assumption is not related to occlusion or nondepiction, but it will play an important role in the characterization theorems.

Note that we cannot make such an assumption about regions, since not all regions need depict a surface, even with no errors in edge detection; in such cases, the region is simply background framed by the edges of overlapping surfaces.

We will see that the occlusion assumption \textit{NLEA} introduced in the previous section is perhaps the most crucial assumption we will make in this work, as it illustrates and makes explicit the relationship between depiction and occlusion. In fact, we will show that this assumption is equivalent to a number of depiction assumptions. For example, consider Figure 8.2a. Here we have a surface in which part of an edge is not depicted because of lacunate errors; as a result the surface is not depicted by a region since a region does not exist. In Figure 8.2b, a region depicting the surface does exist, but the region depicts multiple surfaces. In Figure 8.2c, the region exists and it uniquely depicts the surface; however, part of the region does not depict a surface, but is really part of the background. All of these examples deal with regions and depiction: what exactly is the relationship between these depiction assumptions and the occlusion assumption \textit{NLEA}?

8.1.2 Preconditions for Propositional Theories

One of the primary motivations for investigating NLEA is to provide sufficient conditions for defining sets of propositional theories which are equivalent to \( T_{\text{cardworld}} \cup I \) for some planar image \( I \). Models of \( T_{\text{cardworld}} \cup I \) are not restricted with respect to their domains: although we make closure assumptions concerning the domain of image objects, there are no such restrictions on the domain of scene objects. In Chapter 3, we showed how the Depicted Kernel Assumption can characterize the set of models, allowing us to restrict our attention to surfaces which have some depicted element. However, this places no bound on the number of scene objects that can be depicted by the same image object. For example, it is consistent for a single region to depict an unbounded set of surfaces. We will therefore need to find assumptions which allow us to place a finite bound on the size of the domain for models. In particular, one of the assumptions we will introduce forces regions to depict unique surfaces. Together with the Depicted Kernel Assumption and closure over the image domain, this will allow us to place a finite bound on the domain of scene objects.

The literals that we will be using for the propositional theory are defined as follows:

Definition 8.3
\[
(\forall l, r) \Lambda(l, r) \equiv (\exists e, s) \Delta(l, e) \land \Delta(r, s) \land \text{part}(e, s) \land \text{in}(l, r)
\]
Figure 8.2: Examples of images with models that do not satisfy NLEA.
In other words, $\Lambda(l, r)$ iff line $l$ depicts an edge in the surface depicted by the region $r$ and $l$ is in this region. As a minimal assumption, we will need to define the conditions that guarantee the existence of positive $\Lambda(l, r)$ literals for all lines $l$ and regions $r$.

### 8.2 Regions and Connected Contours

Before introducing the depiction assumptions which are equivalent to $NLEA$, we will prove several results related to $NLEA$ and the existence of cues for occlusion. These results will play important roles in several theorems later in the chapter.

We have seen that the axioms in $T_{occ}$ already guarantee the existence of occluding edges when one surface is on top of another: what they do not guarantee is that these occluding edges will be depicted. If there are lacunary errors in edge detection, these edges may not be depicted.

**Example:** Let $I$ be the image in Figure 8.2d. There is a model $M$ of $T_{cardworld} \cup I$ such that

$$\{e_1, e_2, p_1, s_1, s_2\} \subseteq M$$

and

$$\langle l_1, e_1 \rangle, \langle l_2, e_2 \rangle \in \Delta$$

$$\langle e_1, s_1 \rangle, \langle e_2, s_2 \rangle \in \text{part}$$

$$\langle p_1, e_1 \rangle \in \text{part}, \langle p_1, t_1 \rangle \in \text{position}$$

$$\langle e_1, s_2 \rangle \in \text{occluding}, \langle s_1, s_2, t_1 \rangle \in \text{obscures}$$

However, the occluding edge $e_1$ is not depicted at $t_1$, the position of occlusion.

However, it is not only the occluding edge that must be depicted: the unoccluded points in a partially occluded edge must also be depicted. The image in Figure 8.2(e) violates $NLEA$: in this case, the occluding edges are depicted but there are points in the occluded edge, such as the point that has the position denoted by the constant $t_1$, which are not depicted and not occluded.

Now consider images where there are no lacunary errors in edge detection.

**Example:** Let $I$ be the image in Figure 8.3(a). There is a model $M$ of $T_{cardworld} \cup I$ such that

$$\{e_1, e_2, s_1, s_2\} \subseteq M$$

and

$$\langle r_1, s_1 \rangle, \langle r_2, s_2 \rangle, \langle l_1, e_1 \rangle, \langle l_2, e_2 \rangle, \langle l_3, e_1 \rangle \in \Delta$$

$$\langle e_1, s_1 \rangle, \langle e_2, s_2 \rangle \in \text{part}$$

$$\langle e_1, s_2 \rangle \in \text{occluding}$$

while the image satisfies

$$I \models \text{intersect}(l_1, l_2, j_1) \land \text{position}(j_1, t_1) \land \text{intersect}(l_3, l_2, j_1)$$
Figure 8.3: Hanging lines.
In this case we have what is traditionally referred to in the computer vision literature as a T-junction, since the lines depicting the occluding edge and the occluded edge form a “T”. Depicted occluding edges need not form this kind of T-junction.

Example: Let I be the image in Figure 8.3(b). There is a model \( M \) of \( T_{\text{cardworld}} \cup I \) such that

\[
\{e_1, e_2, s_1, s_2\} \subseteq M \quad \text{and} \quad \langle r_1, s_1 \rangle, \langle r_2, s_2 \rangle, \langle l_1, e_1 \rangle, \langle l_2, e_2 \rangle \in \Delta
\]

\[
\langle e_1, s_1 \rangle, \langle e_2, s_2 \rangle \in \text{part}
\]

\[
\langle e_1, s_2 \rangle, \langle e_2, s_1 \rangle \in \text{occluding}
\]

while the image satisfies

\[
I \models \text{intersect}(l_1, l_2, j_i)
\]

which forms a “V” junction. \( \square \)

What is common to both of these examples is that the occluding and occluded edges are depicted by lines intersecting at the position of occlusion.

In these examples, a point in the occluding edge was depicted by a pixel at the position \( t \). In a sense, the point \( p' \) that is part of the occluding edge can be regarded as the point of occlusion. It is the existence of this depicted occluding point (or rather the existence of depicted occluding edges at the point of occlusion) that is related to the cues used for occlusion, such as T-junctions. For example, in Figure 8.2(e), there is no such point of occlusion since the nondepicted points in the surfaces are not occluded. Since this image violates \( NLEA_4 \), we may ask whether \( NLEA_4 \) guarantees the existence of cues such as T and V junctions.

To enhance readability, we will introduce a new relation with the following conservative definition. We will say that a line is hanging iff a point in the edge depicted by the line is not depicted and there is no depicted occluding edge for the point:

**Definition 8.4** A line is hanging iff it does not intersect two distinct lines

\[
(\forall l) \text{hanging}(l) \equiv \neg(\exists l', j, j') \text{intersect}(l, l', j) \land \text{intersect}(l, l'', j') \land l' \neq l''
\]

Thus, if all points in an edge are depicted, then the line depicting the edge will intersect other lines that depict edges in the same surface. If there is a depicted occluding edge for undepicted points in the edge, then the line depicting the edge will intersect other lines depicting occluding edges. Note that in Figure 8.2d the lines depicting the occluding and occluded edges are hanging lines and in Figure 8.2e the occluded edge is depicted by a hanging line.

The intuition for hanging lines is closely related to axiom 7.18 of \( T_{\text{occ}} \), which is equivalent to saying that hanging lines can only result from undepicted points in edges, whether they are edges in the same surface, abutting edges, or occluding edges. It is reasonable, therefore, to assume that hanging lines cannot exist if there are no errors in edge detection.
Definition 8.5 The Connected Contour Assumption (CCA) is the sentence

\[(\forall l)(\exists l', l'', j, j') \text{ intersect}(l, l', j) \land \text{intersect}(l, l'', j') \land l' \neq l''\]

Recalling the definition of hanging lines, CCA can also be expressed as

\[(\forall l) \neg\text{hanging}(l)\]

The sentence CCA is an image sentence, and is not in itself related to depiction. In the following assumption we restrict CCA to lines that depict edges:

Definition 8.6 The Weak Connected Contour Assumption (WCCA) is the sentence

\[(\forall l, e) \Delta(l, e) \supset (\exists l', l'', j, j') \text{ intersect}(l, l', j) \land \text{intersect}(l, l'', j') \land l' \neq l''\]

Similarly, WCCA can be rewritten as

\[(\forall l, e) \Delta(l, e) \supset \neg\text{hanging}(l)\]

Throughout this work we will be assuming NSEA, so that there are no spurious errors in edge detection. In this case, the equivalence between CCA and WCCA follows easily from the definition of NSEA:

Proposition 8.1

\[T_{\text{cardworld}} \cup \text{NSEA} \models \text{CCA} \equiv \text{WCCA}\]

The following lemma defines the relationship between occlusion, depiction and the existence of intersecting lines.

Lemma 8.1

\[T_{\text{cardworld}} \models \text{NLEA} \supset \text{WCCA}\]

Proof: Suppose that the theorem is false:

\[T_{\text{cardworld}} \not\models \text{NLEA} \supset \text{WCCA}\]

Then there exists a model \(\mathcal{M}\) of \(T_{\text{cardworld}} \cup \text{NLEA}\) such that

\[\mathcal{M} \models (\exists l, e) \Delta(l, e) \land \text{hanging}(l)\]

By axiom 4.2 of \(T_{\text{fg}}\) we have

\[\mathcal{M} \models (\exists l, e, e', v, s) \Delta(l, e) \land \text{hanging}(l) \land \text{part}(e, s) \land \text{part}(e', s) \land \text{meet}(e, e', v)\]

By axiom 7.18 of \(T_{\text{occ}}\) we have

\[\mathcal{M} \models (\exists l, e, e', p, t_i, s) \Delta(l, e) \land \text{hanging}(l) \land \text{part}(e, s) \land \text{part}(e', s) \land \text{meet}(p, e', s)\]

By \(\text{NLEA}\) and axiom 7.15 of \(T_{\text{occ}}\) we have

\[\mathcal{M} \models (\exists l, e, e', p, t_i, s, s') \Delta(l, e) \land \text{hanging}(l) \land \text{part}(e, s) \land \text{part}(e', s)\]
Figure 8.4: Phase 1 in the characterization of assumptions about uniquely depicting regions. Lines in the diagram represent entailment among the assumptions.

\[ \land (\text{part}(p, e) \lor \text{part}(p, e')) \land \text{part}(p, s) \land \text{position}(p, t_1) \land \neg (\exists q) \Delta(q, p) \land \text{obscures}(s', s, t_1) \]

By axiom 7.10 of \( T_{\text{obsures}} \), we restrict ourselves to the minimal obscuring surface. By the existence of occluding edges and axiom 7.8 of \( T_{\text{occ}} \):

\[ \forall l \models (\exists e, e', p, t_1, s, s') \Delta(l, e) \land \text{hanging}(l) \land \text{part}(e, s) \land \text{part}(e', s) \]

\[ \land (\text{part}(p, e) \lor \text{part}(p, e')) \land \text{part}(p, s) \land \text{position}(p, t_1) \land \neg (\exists q) \Delta(q, p) \land \text{obscures}(s', s, t_1) \land \text{part}(e', s') \land \text{occluding}(e', s) \]

The definition of occluding gives

\[ \forall l \models (\exists p', t_2, l, e, e', p, t_1, s, s') \Delta(l, e) \land \text{hanging}(l) \land \text{part}(e, s) \land \text{part}(e', s) \]

\[ \land (\text{part}(p, e) \lor \text{part}(p, e')) \land \text{part}(p, s) \land \text{position}(p, t_1) \land \neg (\exists q) \Delta(q, p) \land \text{obscures}(s', s, t_1) \land \text{part}(p', e) \land \text{part}(e', s') \land \text{position}(p', t_2) \land \text{occluding}(e', s) \land \text{obscures}(s', s, t_2) \]

Since this point is not occluded, by \( NLEA \) it must be depicted, so that we get

\[ \models (\exists q', p', l, e, e', p, t_1, s, s', q') \Delta(l, e) \land \text{hanging}(l) \land \text{part}(e, s) \]

\[ \land \text{part}(e', s) \land (\text{part}(p, e) \lor \text{part}(p, e')) \land \text{part}(p, s) \land \text{position}(p, t_1) \land \neg (\exists q) \Delta(q, p) \land \text{obscures}(s', s, t_1) \land \text{part}(p', e) \land \text{part}(e', s') \land \text{position}(p', t_2) \land \text{occluding}(e', s) \land \text{obscures}(s', s, t_2) \land \Delta(q', p') \]

which leads to a contradiction by axiom 7.18 of \( T_{\text{occ}} \). □

We can again notice that axiom 7.8 plays a crucial role in the proof. This axiom entails the existence of occluding edges: if such edges exist, then they will be depicted as intersecting lines.

8.3 Assumptions about Depicting Regions

In this section, we explore the relationship between \( NLEA \) and the assumption that regions depict unique surfaces. This assumption of uniquely depicting regions is a crucial property - if there is no bound on the number of surfaces depicted by a region, there is no bound to the size of the domain of a model, and we must allow infinite models even with the Depicted Kernel Assumption. We therefore introduce an assumption that imposes a bound on the number of depicted surfaces; it will later allow us to compute propositional models of an image and the axioms.
Regions depict unique surfaces (URDA)

Every position included in a region that depicts a surface is also interior to that surface (INDA)

Every line in a region that depicts an edge depicts either an occluding edge or an abutting edge or an edge in some surface depicted by the region (WDTR)

If a position is interior to a hole of a surface and included in a region, then the region that does not depict the surface (HREA)

Figure 8.5: Phase 2 in the characterization of assumptions about uniquely depicting regions. Lines in the diagram represent entailment among the assumptions.
There are two phases to this section, illustrated by Figures 8.4 and Figure 8.5. In these figures, each box corresponds to a different depiction assumption that we are introducing, together with the acronym used to name the assumption in the theorems.

In the first phase, we introduce the assumption that every position included in a region that depicts a surface is also interior to that surface. This assumption is weaker than NLEA, but it is strong enough to entail the assumption about uniquely depicting regions. In the second phase, we show that this new assumption is actually equivalent to the conjunction of the following three assumptions:

- Regions depict unique surfaces.
- Every line in a region depicts either an occluding edge or an abutting edge.
- Every position interior to a hole in a surface is not included in a region that depicts the surface.

### 8.3.1 Uniquely Depicting Regions

In models of $T_{\text{cardworld}} \cup I$ which are inconsistent with $NLEA$, such as images with silhouettes of sets of surfaces in Figures 8.2(b) and 8.2(e), regions can depict multiple surfaces. We will first introduce the assumption that forces regions to depict unique surfaces, and show that this assumption is entailed by $NLEA$. This will lead us to the derivation of several other assumptions related to the depiction of interior and inclusion.

**Definition 8.7** The Unique Region Depiction Assumption (URDA) is the sentence

$$(\forall r, s, s') \text{region}(r) \land \Delta(r, s) \land \Delta(r, s') \supset s = s'$$

i.e. every region depicts a unique surface.

One of the characterization theorems in this section will determine the depiction and occlusion assumptions necessary and sufficient to entail $URDA$. We will prove that $URDA$ is entailed by $NLEA$ in two steps - we will first show that $NLEA$ entails a stronger assumption, and then show that this assumption entails $URDA$.

Let us take a closer look at images with models that violate $URDA$ to uncover what this stronger assumption could be. Consider Figure 8.2b, which has a model that violates $URDA$.

**Example:** Let $I$ be the image in Figure 8.2(b). For this image, which satisfies

$I \models in(l_1, r_1) \land in(l_2, r_1) \land inclusion(r_1, t_1) \land inclusion(r_1, t_2)$

there is a model $M$ of $T_{\text{cardworld}} \cup I$ such that $\{e_1, e_2, s_1, s_2\} \subseteq M$ and

$$\langle r_1, s_1 \rangle, \langle l_1, e_1 \rangle, \langle l_1, e_2 \rangle \in \Delta$$

$$\langle e_1, s_1 \rangle, \langle e_2, s_2 \rangle \in \text{part}$$
Note, however, that

\( \langle s_1, t_1 \rangle, \langle s_2, t_2 \rangle \in \text{interior} \)

even though the position is included in the region depicting \( s_1 \) and

\( \langle s_1, t_2 \rangle \notin \text{interior} \)

even though the position is included in the region depicting \( s_2 \). □

Consider Figure 8.2c, which has a model that satisfies \( URDA \), but which does not satisfy \( NLEA \):

Example: Let \( I \) be the image in Figure 8.2(c). For this image, which satisfies

\[ I \models inclusion(r_1, t_1) \land inclusion(r_1, t_2) \]

there is a model \( M \) of \( T_{\text{cardworld}} \cup I \) such that \( \{ s_1 \} \subseteq M \) and

\[ \langle r_1, s_1 \rangle \in \Delta \]

\( \langle s_1, t_1 \rangle \in \text{interior}, \langle s_1, t_2 \rangle \notin \text{interior} \)

so that there is a position that is not interior to a surface even though the position is included in the region depicting \( s_1 \). □

To eliminate these cases, we introduce the following assumption.

Definition 8.8 The Inclusion Depiction Assumption (\( INDA \)) is the sentence

\[ (\forall r, s, t) \; \Delta(r, s) \land inclusion(r, t) \supset \text{interior}(s, t) \]

i.e. every position included in a region that depicts a surface is also interior to that surface.

In all of the above examples of models that violate \( INDA \), we can see that the the boundary of a surface is included in the region depicting the surface. Intuitively, \( INDA \) should prevent this from happening. We can express this intuition as the following consequence of \( INDA \), which will be useful in later proofs:

Lemma 8.2 For any planar image \( I \),

\[ T_{\text{cardworld}} \cup I \cup INDA \models (\forall r, s, t) \; \text{inclusion}(r, t) \land \Delta(r, s) \supset (\exists p) \; \text{part}(p, s) \land \text{position}(p, t) \]

i.e., if we assume \( INDA \), then the boundary of a surface cannot be included in the region depicting the surface.

Proof: By \( INDA \), a position which is interior to the surface must be included in the region that depicts the surface. Axiom 6.12 of \( T^\Delta_{\text{interior}} \) then entails that all depicted points in the surface be contained in a border line of the region. By the definition of planar images, the positions of such points cannot be included in the region. □
Notice that the two examples whose models violate \textit{INDA}, also violate \textit{NLEA}. The following theorem\footnote{Any theorem whose proof requires that we reason about the intended interpretations of predicates in \(L_{\text{planar}}\), will be restricted to planar images.} shows that indeed \textit{INDA} is a consequence of the assumption that there are no lacunate errors in edge detection.

**Theorem 8.1** For any planar image \(I\),

\[
T_{\text{cardworld}} \cup I \models NLEA \supset INDAA
\]

**Proof:** We will use Lemma 8.2, and suppose that it is falsified, so that

\[
T_{\text{cardworld}} \cup I \cup NLEA \models (\exists r, s, e, p, t) \Delta(r, s) \land \text{inclusion}(r, t) \land \text{part}(e, s)
\]

\[
\land \text{part}(p, e) \land \text{position}(p, t)
\]

\[
\land (\exists s') \text{obscures}(s', s, t)
\]

We will prove the theorem by cases. In the first case, we will assume that the surface depicted by the region is not obscured by any other surface. In the second case, we will assume that the surface depicted by the region is obscured by another surface at some position.

- **Case 1:**

\[
T_{\text{cardworld}} \cup I \cup NLEA \models (\exists r, s, e, p, t) \Delta(r, s) \land \text{inclusion}(r, t) \land \text{part}(e, s)
\]

\[
\land \text{part}(p, e) \land \text{position}(p, t)
\]

\[
\land (\exists s') \text{obscures}(s', s, t)
\]

Since the point \(p\) is not occluded, by \textit{NLEA} it will be depicted. Since \textit{NLEA} entails \textit{WCCA}, the pixel depicting the point will be contained in a non-hanging line:

\[
T_{\text{cardworld}} \cup I \cup NLEA \models (\exists r, s, e, p, t) \Delta(r, s) \land \text{inclusion}(r, t) \land \text{part}(e, s)
\]

\[
\land \text{part}(p, e) \land \text{position}(p, t)
\]

\[
\land (\neg(\exists s') \text{obscures}(s', s, t)) \land (\exists q) \Delta(l, e) \land \neg\text{hanging}(l) \land \text{in}(q, l) \land \Delta(q, p) \land \text{position}(q, t)
\]

However, this says that there exist pixels which are included in the region and which are also contained in a line depicting an edge in the surface, which is inconsistent with the existence of border lines in a planar image.

- **Case 2:**

\[
T_{\text{cardworld}} \cup I \cup NLEA \models (\exists r, s, e, p, t) \Delta(r, s) \land \text{inclusion}(r, t) \land \text{part}(e, s)
\]

\[
\land \text{part}(p, e) \land \text{position}(p, t) \land (\exists s') \text{obscures}(s', s, t)
\]

By axiom 7.10 of \(T_{\text{obscures}}\), we can restrict ourselves to the minimal obscuring surface. By the Occluding Edge Existence Lemma, there exists an occluding edge that is part of the obscuring surface \(s'\):

\[
T_{\text{cardworld}} \cup I \cup NLEA \models (\exists r, s, e, p, t) \Delta(r, s) \land \text{inclusion}(r, t) \land \text{part}(e, s)
\]

\[
\land \text{part}(p, e) \land \text{position}(p, t) \land (\exists s') \text{obscures}(s', s, t)
\]

\[
\land (\exists s') \text{part}(e', s') \land \text{occluding}(e', s)
\]

Since \textit{NLEA} entails \textit{WCCA}, the pixel depicting the point in the occluding edge will be contained in a non-hanging line:

\[
T_{\text{cardworld}} \cup I \cup NLEA \models (\exists r, s, e, p, t) \Delta(r, s) \land \text{inclusion}(r, t) \land \text{part}(e, s)
\]
\[ \begin{align*} 
& \land \text{part}(p, e) \land \text{position}(p, t) \land (\exists s') \text{ obscures}(s', s, t) \\
& \land (\exists t_2) \text{ part}(e, s') \land \text{occluding}(e', s) \land \text{inclusion}(r, t_2) \\
& \land (\exists l, t) \text{ in}(q, l) \land \Delta(l, e') \land \neg \text{hanging}(l) \land \text{position}(q, t_2) 
\end{align*} \]

However, this says that there exist pixels which are included in the region and which are also contained in a line depicting an occluding edge in the surface, which is inconsistent with the existence of border lines in a planar image.

\[ \square \]

We can also show that \( URDA \) is a consequence of \( INDA \); in conjunction with the previous theorem, this enables us to prove that \( URDA \) is also entailed by the assumption of no lacunate errors in edge detection.

**Theorem 8.2** For any planar image \( I \),

\[
T_{\text{cardworld}} \cup I \models INDA \supset URDA
\]

**Proof:** Suppose \( URDA \) is falsified, that is,

\[
T_{\text{cardworld}} \cup I \cup INDA \models (\exists r, s, s') \Delta(r, s) \land \Delta(r, s') \land s \neq s'
\]

Then by axiom 7.17 of \( T_{\text{obscura}} \) (the occlusion axiom for surfaces) there exist positions included in the region at which neither surface occludes the other:

\[
T_{\text{cardworld}} \cup I \cup INDA \models (\exists r, s, s', t_1, t_2) \Delta(r, s) \land \Delta(r, s') \land s \neq s' \\
\land \neg \text{obscures}(s, s', t_1) \land \neg \text{obscures}(s', s, t_1) \\
\land \text{inclusion}(r, t_1) \land \text{inclusion}(r, t_2) \land \text{interior}(s, t_1) \land \text{interior}(s', t_2)
\]

There are two cases, depending on whether or not these positions are interior to both surfaces.

**Case 1:**

\[
T_{\text{cardworld}} \cup I \cup INDA \models (\exists r, s, s', t_1, t_2) \Delta(r, s) \land \Delta(r, s') \land s \neq s' \land (\text{interior}(s, t_2) \setminus \text{interior}(s', t_1))
\]

By axiom 7.1 of \( T_{\text{obscura}} \), we have

\[
T_{\text{cardworld}} \cup I \cup INDA \models (\exists r, s, s', t_1, t_2) \Delta(r, s) \land \Delta(r, s') \land s \neq s' \land \neg \text{obscures}(s, s', t_2) \land \neg \text{obscures}(s', s, t_1)
\]

By the Occluding Edge Existence Lemma, there exists an occluding edge, since both surfaces are only partially occluded:

\[
T_{\text{cardworld}} \cup I \cup INDA \models (\exists r, s, s', t_1, t_2) \Delta(r, s) \land \Delta(r, s') \land s \neq s' \land \neg \text{obscures}(s', s, t_2) \land \neg \text{obscures}(s, s', t_1) \\
\land (\exists e, p, t_3) \text{part}(e, s) \land \text{occluding}(e, s') \land \text{part}(p, e) \land \text{position}(t_3) \land \text{inclusion}(r, t_3)
\]

which falsifies Lemma 8.2.

**Case 2:**

\[
T_{\text{cardworld}} \cup I \cup INDA \models (\exists r, s, s', t_1, t_2) \Delta(r, s) \land \Delta(r, s') \land s \neq s' \land \neg \text{interior}(s, t_2) \land \neg \text{interior}(s', t_1)
\]

By axiom 6.7 of \( T_{\text{obscura}} \), there exists a boundary edge for the surface \( s \) such that

\[
T_{\text{cardworld}} \cup I \cup INDA \models (\exists r, s, s', t_1, t_2) \Delta(r, s) \land \Delta(r, s') \land s \neq s' \\
\land (\exists e, p, t_3) \text{boundary}(e, s) \land \text{part}(p, e) \land \text{position}(t_3) \land \text{inclusion}(r, t_3)
\]

which falsifies Lemma 8.2. \( \square \)
Corollary 8.1 For any planar image $I$, 

$$T_{\text{cardworld}} \cup I \models NLEA \supset URDA$$

It should be noted that a critical role is played by the Occluding Edge Existence Lemma in both of these proofs. In the proof of $NLEA \supset INDA$, these edges must be depicted; with $INDA \supset URDA$, the position of points in the occluding edge cannot be included in the region depicting the surface. In each case, $NLEA$ merely guarantees that unoccluded points are depicted: the occlusion axioms guarantee that the boundary edges for positions interior to the surfaces are not occluded.

### 8.3.2 The Depicted Trichotomy Assumption

At the beginning of this section we stated that we would decompose the occlusion assumption $NLEA$ into a set of independent depiction assumptions. At this point we have the depiction assumption $INDA$ which is entailed by $NLEA$. We will now proceed in a similar manner in decomposing $INDA$ into a set of independent depiction assumptions, which is Phase 2 in this section (see Figure 8.5). As the image in Figure 8.2(c) illustrates, it is possible to satisfy $URDA$ and still violate $INDA$, since a region may depict a unique surface even though part of the region does not depict that surface. If we wish to have a complete characterization of $INDA$, we need to find other assumptions that are entailed by $INDA$ and which are independent of $URDA$. To uncover these new assumptions, we will take a closer look at images like Figure 8.2(c).

One intuitive consequence of $INDA$ is that the border of the region depicting a surface coincides with the boundary edges of the surface unless an edge of the surface is occluded by another surface. For example, in Figure 8.2(c), there exist border lines of the region $r_i$ that depict neither edges in the surface $s_1$, nor occluding edges for $s_1$.

**Example:** Consider the image $I$ in Figure 8.6(a) which satisfies

$$I \models in(l_1, r_1) \land in(l_2, r_1) \land in(l_3, r_1) \land in(l_4, r_1) \land in(l_5, r_1) \land in(l_3, r_2)$$

There is a model $M$ of $T_{\text{cardworld}} \cup I$ such that \{e_1, e_2, e_3, e_4, e_5, s_1, s_2, s_3\} $\subset$ $M$ and that satisfies $NLEA$ such that

$$\langle r_1, s_1 \rangle, \langle r_2, s_2 \rangle, \langle r_3, s_3 \rangle, \langle l_1, e_1 \rangle, \langle l_2, e_2 \rangle, \langle l_3, e_2 \rangle, \langle l_4, e_4 \rangle, \langle l_5, e_5 \rangle, \langle l_3, e_6 \rangle \in \Delta$$

$$\langle e_1, s_1 \rangle, \langle e_2, s_1 \rangle, \langle e_3, s_1 \rangle, \langle e_4, s_1 \rangle, \langle e_5, s_3 \rangle, \langle e_6, s_2 \rangle \in \text{part}$$

In this model, $e_5$ denotes an occluding edge in the surface $s_3$ and the edges $e_3$ and $e_6$ denote abutting edges. $\square$

We encapsulate these intuitions in the following assumption:
Figure 8.6: Examples of DTR and HREA
Definition 8.9 The Depicted Trichotomy Assumption (DTR) is the sentence

\[(\forall r, s, l, e) \Delta(r, s) \land \text{in}(l, r) \land \Delta(l, e) \supset \text{occluding}(e, s, t) \lor (\exists e'), t \text{ abuts}(e, e', t) \lor \text{part}(e, s)\]

i.e. every line in a region that depicts an edge in the surface depicted by the region depicts either an occluding edge or an abutting edge.

The following theorem shows that this new depiction assumption is entailed by \text{INDA}:

Theorem 8.3 For any planar image \(I\),

\[T_{\text{cardworld}} \cup I \models \text{INDA} \supset DTR\]

Proof: Since the following sentence is entailed by any planar image \(I\),

\[I \models (\forall l, q, r, t) \text{in}(l, r) \land \text{in}(q, l) \land \text{position}(q, t) \supset \neg \text{exclusion}(r, t)\]

\text{INDA} entails

\[T_{\text{cardworld}} \cup I \cup \text{INDA} \models (\forall l, q, r, s, t) \text{in}(l, r) \land \text{in}(q, l) \land \text{position}(q, t) \land \Delta(r, s) \supset \text{interior}(s, t)\]

that is, if we assume \text{INDA}, then the position of any pixel contained in a region depicting some surface is interior to that surface.

Since all edges must be part of some surface,

\[T_{\text{cardworld}} \cup I \cup \text{INDA} \models (\forall l, q, r, s, e, t) \text{in}(l, r) \land \text{in}(q, l) \land \text{position}(q, t) \land \Delta(r, s) \supset \text{interior}(s, t)\]

\[\land (\neg \text{part}(e, s) \lor (\exists s') \text{part}(e, s') \land \text{interior}(s', t))\]

so that

\[T_{\text{cardworld}} \cup I \models (\forall l, q, r, s, e, t) \text{in}(l, r) \land \text{in}(q, l) \land \text{position}(q, t) \land \Delta(r, s) \supset \text{interior}(s, t)\]

\[\land (\exists s') \text{interior}(s', t) \lor \text{part}(e, s)\]

Axioms 7.1, 7.14, and 7.8 of \(T_{\text{ooc}}\) then give us

\[T_{\text{cardworld}} \cup I \models (\forall l, q, r, s, t) \text{in}(l, r) \land \text{in}(q, l) \land \text{position}(q, t) \land \Delta(r, s) \supset \text{interior}(s, t)\]

\[\land \text{occluding}(e, s) \lor \text{part}(e, s) \lor (\exists e') \text{abuts}(e, e', t)\]

\(\Box\)

Depicted Trichotomy is a rather strong assumption, as we can see from the following examples.

Example: Consider the image \(I\) in Figure 8.3(c). There is a model \(M\) of \(T_{\text{cardworld}} \cup I\) such that

\[\{e_1, s_1, s_2, s_3\} \subset M\] and

\[\langle r_1, s_1 \rangle, \langle r_1, s_2 \rangle, \langle r_2, s_3 \rangle, \langle l_1, e_1 \rangle \in \Delta\]

\[\langle e_1, s_3 \rangle \in \text{part}, \langle e_1, s_2 \rangle \in \text{occluding}\]

However, this model violates \(DTR\) since

\[\langle e_1, s_1 \rangle \notin \text{occluding}\]

and there does not exist an edge \(e\) such that

\[\langle e, e_1, t \rangle \in \text{abuts}\]

Figure 8.3(d) illustrates the scene with all undepicted edges displayed. \(\Box\)
The reason that the image in this example violates DTR is that it also violates U'RA. Even though every line in the region depicts an occluding or abutting edge in some surface, DTR is violated because there are multiple surfaces depicted by the region and there exists a surface depicted by the region which is not occluded or abutting. It seems, therefore, that DTR implicit assumes that regions depict unique surfaces. To handle this problem, we can weaken this assumption as follows:

**Definition 8.10** *The Weak Depicted Trichotomy Assumption (WDT)* is the sentence

\[(\forall r, s, l, e) \Delta(r, s) \land \text{in}(l, r) \land \Delta(l, e) \supset \]

\[(\exists s') \Delta(r, s') \land (\text{occluding}(e, s') \lor (\exists e', t) \text{abuts}(e, e', t) \lor \text{part}(e, s'))\]

i.e. every line in a region that depicts an edge in the surface depicted by the region depicts either an occluding edge, an abutting edge, or an edge in some surface depicted by the region.

This assumption handles multiply depicting regions such as those in Figure 8.3c, since we only require that some surface depicted by the region be occluded or abutting.

It is easy to see from the definitions that WDT is weaker than DTR:

**Proposition 8.2**

\[T_{\text{cardworld}} \models DTR \supset WDT\]

However, if we assume that regions depict unique surfaces, then WDT and DTR are equivalent:

**Proposition 8.3**

\[T_{\text{cardworld}} \models U'RA \supset (WDT \equiv DTR)\]

The major result of this section follows easily from the definitions and Theorem 8.3:

**Corollary 8.2** *For any planar image I,*

\[T_{\text{cardworld}} \cup I \models INDA \supset WDT\]

Thus, we have found another assumption which is both entailed by INDA and which is independent of U'RA.

**8.3.3 The Hole Region Existence Assumption**

Before presenting our final assumption in the characterization of INDA, notice that even with isolated surfaces, regions do not always depict surfaces. For instance, if the surface has a hole in it, as in Figure 8.6(b) then the hole will also be "depicted" by a region, even though the hole is not part of the figure but rather part of the background. This fact can lead to images that satisfy U'RA \land WDT yet violate INDA.

**Example:** Consider Figure 8.6(c) in which the image I satisfies

\[I \models \text{in}(l_1, r_1) \land \text{inclusion}(r_1, t_1) \land \text{inclusion}(r_1, t_2) \land \text{in}(l_2, r_1)\]
There is a model $\mathcal{M}$ of $T_{\text{card\_world}} \cup I$ such that \( \{e_1, s_1\} \subseteq M \) and
\[
< r_1, s_1 >, < l_1, e_1 > \in \Delta, < e_1, s_1 > \in \text{part}
\]
\[
< s_1, t_2 > \in \text{interior}, < s_1, t_1 > \notin \text{interior}, < s_1, t_1 > \in \text{hole\_interior}
\]

Note that this model satisfies both $URDA$ and $WDTR$. $\square$

Using the axioms about holes and their depiction from $T_{\text{interior}}$ and $T_{\text{interior}}^\Delta$, we are now ready to state our final assumption in the characterization of $INDA$:

**Definition 8.11** The Hole Region Existence Assumption ($HREA$) is the sentence
\[
(\forall r, s, t) \ \text{hole\_interior}(s, t) \land \text{inclusion}(r, t) \supset -\Delta(r, s)
\]
i.e., if a position is interior to the hole of a surface and included in a region, then the region does not depict the surface.

Note that the preceding example violates this assumption, since the position $t_2$ is interior to the hole in the surface $s_1$, yet it is also included in the region $r_1$, which depicts $s_1$.

The following theorem shows that $HREA$ is entailed by $INDA$:

**Theorem 8.4**

\[
T_{\text{card\_world}} \models INDA \supset HREA
\]

**Proof:** Suppose $HREA$ is falsified:

\[
T_{\text{card\_world}} \cup I \models (\exists r, s, t) \ \text{hole\_interior}(s, t) \land \text{inclusion}(r, t) \land \Delta(r, s)
\]

By the definition of $\text{hole\_interior}$ from $T_{\text{interior}}$, a position cannot be both interior to a surface and also interior to a hole of a surface, so we must have

\[
T_{\text{card\_world}} \cup I \models (\exists r, s, t) \ \text{inclusion}(r, t) \land \text{non-interior}(s, t) \land \Delta(r, s)
\]

which is equivalent to $-INDA$. $\square$

We can now give a complete characterization of the assumption $INDA$:

**Theorem 8.5** For any planar image $I$,

\[
T_{\text{card\_world}} \cup I \models URDA \land WDTR \land HREA \supset INDA
\]

**Proof:** We will show

\[
T_{\text{card\_world}} \cup I \models -INDA \land URDA \land HREA \supset -WDTR
\]

Suppose, then, that $INDA$ is falsified so that:

\[
T_{\text{card\_world}} \cup I \land -INDA \land URDA \land HREA \models (\exists r, s, t) \ \Delta(r, s) \land \text{inclusion}(r, t) \land \text{non-interior}(s, t)
\]

By Lemma 8.2, there must exist a point in the surface whose position is included in the region, so that we have

\[
T_{\text{card\_world}} \cup I \land -INDA \land URDA \land HREA \models (\exists r, s, t_1, p, t_2) \ \Delta(r, s) \land \text{inclusion}(r, t_1)
\]
\[ \neg \text{interior}(s, t_1) \land \text{part}(p, s) \land \text{position}(p, t_2) \land \text{inclusion}(r, t_2) \]

Axiom 6.14 of \( T_{interior} \) entails the existence of a non-outer (hole) edge containing the point:

\[ T_{cardworld} \cup I \land \neg \text{INDA} \land URDA \land HREA \models (\exists r, s, e, t_1, p, t_2) \Delta(r, s) \land \text{inclusion}(r, t_1) \land \neg \text{interior}(s, t_1) \land \text{part}(p, e) \land \text{position}(p, t_2) \land \text{inclusion}(r, t_2) \land \neg \text{outer}(e) \land \text{hole.boundary}(e, t_2) \]

\[ HREA \] and axiom 6.14 of \( T^A_{interior} \) entails that the position is not interior to any hole of the surface \( s \):

\[ T_{cardworld} \cup I \land \neg \text{INDA} \land URDA \land HREA \models (\exists r, s, e, t_1, p, t_2) \Delta(r, s) \land \text{inclusion}(r, t_1) \land \neg \text{interior}(s, t_1) \land \text{part}(p, e) \land \text{position}(p, t_2) \land \text{inclusion}(r, t_2) \land \neg \text{outer}(e) \land \text{hole.boundary}(e, t_2) \land (\exists p', q, l, t_3) \text{in}(l, r) \land \Delta(l, e) \land \Delta(q, p') \land \text{part}(p', e) \land \text{position}(p', t_3) \land \neg \text{hole.interior}(s, t_3) \]

Then \( URDA \) guarantees that the position cannot be interior to any other surface (since only one surface is depicted by the region):

\[ T_{cardworld} \cup I \land \neg \text{INDA} \land URDA \land HREA \models (\exists r, s, e, t_1, p, t_2) \Delta(r, s) \land \text{inclusion}(r, t_1) \land \neg \text{interior}(s, t_1) \land \text{part}(p, e) \land \text{position}(p, t_2) \land \text{inclusion}(r, t_2) \land \neg \text{outer}(e) \land \text{hole.boundary}(e, t_2) \land (\exists p', q, l, t_3) \text{in}(l, r) \land \Delta(l, e) \land \Delta(q, p') \land \text{part}(p', e) \land \text{position}(p', t_3) \land \neg \text{hole.interior}(s, t_3) \land (\exists s') \neg \text{interior}(s', t_3) \]

Using axiom 7.1 of \( T_{occ} \) we get

\[ T_{cardworld} \cup I \land \neg \text{INDA} \land URDA \land HREA \models (\exists r, s, e, t_1, p, t_2) \Delta(r, s) \land \text{inclusion}(r, t_1) \land \neg \text{interior}(s, t_1) \land \text{part}(p, e) \land \text{position}(p, t_2) \land \text{inclusion}(r, t_2) \land \neg \text{outer}(e) \land \text{hole.boundary}(e, t_2) \land (\exists p', q, l, t_3) \text{in}(l, r) \land \Delta(l, e) \land \Delta(q, p') \land \text{part}(p', e) \land \text{position}(p', t_3) \land \neg \text{hole.interior}(s, t_3) \land (\exists s') \neg \text{interior}(s', t_3) \land (\exists l', e') \text{in}(l', r) \land \Delta(l', e') \land \neg \text{part}(e', s) \land \neg \text{occluding}(e', s) \land (\exists e', w) \text{abuts}(e, e', w) \]

which is equivalent to \( \neg WDTR \). \( \square \)

We thus have a complete characterization of the depiction assumption \( IND A \) – we have found three independent depiction assumptions that together are logically equivalent to \( IND A \).

8.3.4 Summary and Implications

In effect, the depiction assumptions we have investigated in this section have established the equivalence between \text{interior} as a scene relation and \text{inclusion} as an image relation when we assume \( IND A \). The scene properties enforced by the axioms of \( T_{occ} \) are preserved in the image by the depiction assumptions \( IND A, URDA, WDTR, \) and \( HREA \). Regions depict unique surfaces; we can forget about the distinction among positions included in the region, since they are all interior to the same surface. All lines in a region depict either edges in the surface or they are occluding edges, so that we can forget about the distinction between boundaries of surfaces and borders of regions.
8.4 Assumptions of Region Existence

In images satisfying \( NLEA \) we notice that regions always exist – lines are always in regions and points interior to a surface are included in some region. On the other hand, these properties do not hold in images where \( NLEA \) is violated, as we notice in Figure 8.1. In this section we introduce the assumptions that guarantee the existence of regions, and show that this assumption is entailed by \( NLEA \). As in the previous section, this will lead us to the derivation of several other depiction assumptions. Assumptions about region existence are essentially specifying the necessary and sufficient conditions for the depiction of surfaces, just as \( NLEA \) specifies the necessary and sufficient conditions for the depiction of points.

There are three phases to this part of the chapter, illustrated by Figures 8.7, 8.8, and 8.9. In these figures, each box corresponds to a different depiction assumption that we are introducing, together with the acronym used to name the assumption in the theorems.

In the first phase (Figure 8.7), we introduce an assumption which states that every depicted edge in a surface is contained in a region that depicts the surface. We show that this assumption is equivalent to three independent depiction assumptions:

- Every line that depicts an edge is contained in a region.
Every position interior to a surface is either included in a region, or it is the position of a pixel depicting a point in the surface.

(BREA)

If an edge in a surface is depicted by a line, then there exists a region containing the line and depicting the surface

(IDA)

Every depicted occluding edge is in a region that depicts either the occluding surface or the occluded surface

(OSDA)

Figure 8.8: Phase 2 in the characterization of assumptions about region existence. Lines in the diagram represent entailment among the assumptions.

- Every surface with a depicted edge is itself depicted by a region.

- If a line depicts an edge, but is contained in a region that does not depict the surface containing the edge, then the line is contained in two distinct regions.

However, we discover that this assumption alone is inadequate to characterize the relationship to NLEA. In particular, we must not only make assumptions about lines being contained in regions, we must also make the assumption that positions interior to a surface is either included in a region, or it is the position of a depicted point of the surface. In the second phase (Figure 8.8), show that this new assumption is equivalent to the conjunction of the following two independent assumptions:

- Every depicted edge in a surface is contained in a region that depicts the surface.

- Every depicted occluding edge is in a region that depicts either the occluding surface or the occluded surface.

Finally, in the third phase (Figure 8.9), we show that this stronger assumption is itself entailed by NLEA, and also that together with IND A (from the preceding section), they are equivalent to NLEA.

### 8.4.1 Existence of Depicting Regions

Earlier in this chapter, we introduced the literals that we will be using for the propositional theories:

\[(\forall l, r) \Lambda(l, r) \equiv (\exists e, s) \Delta(l, e) \land \Delta(r, s) \land \text{part}(e, s) \land \text{in}(l, r)\]
Every position interior to a surface is either included in a region, or it is the position of a pixel depicting a point in the surface. 

NLEA

Every position included in a region that depicts a surface is also interior to that surface

(INDA)

Figure 8.9: Phase 3 in the characterization of assumptions about region existence. Lines in the diagram represent entailment among the assumptions.

In other words, \( \Lambda(l, r) \) iff line \( l \) depicts an edge in the surface depicted by the region \( r \) and \( l \) is in this region. As a minimal assumption, we will need to define the conditions that guarantee the existence of positive \( \Lambda(l, r) \) literals for all lines \( l \) and regions \( r \).

**Definition 8.12** The Depiction Assumption for the image predicate in (IDA) is the sentence

\[
(\forall l, e, s) \text{ part}(e, s) \land \Delta(l, e) \supset (\exists r) \Delta(r, s) \land \text{in}(l, r)
\]

i.e., if a line depicts an edge in some surface, then the line is contained in a region depicting that surface.

The next theorem shows that models of \( T_{\text{cardworld}} \) which satisfy IDA do indeed guarantee that the theory contains positive \( \Lambda \) literals.

**Theorem 8.6** Let \( M \) be a model of \( T_{\text{cardworld}} \cup \text{IDA} \cup \text{NSEA} \); then for every line \( l \in M \) there exists a region \( r \in M \) such that \( (l, r) \in \Lambda \).

**Proof:** By NSEA and \( T_{\text{kernel}} \), there exists an edge \( e \in M \) and a surface \( s \in M \) such that

\[
\langle l, e \rangle \in \Delta, \langle e, s \rangle \in \text{part}
\]

By IDA, there exists a region \( r \in M \) such that

\[
\langle r, s \rangle \in \Delta, \langle l, r \rangle \in \text{in}
\]

so that by the definition of \( \Lambda \) we have

\[
\langle l, r \rangle \in \Lambda
\]

\( \square \)

It is important to see the role that NSEA plays in the existence of positive \( \Lambda \) literals. If a line does not depict an edge (thus violating NSEA), then there is no surface; IDA is satisfied in this case, but there can be no positive \( \Lambda \) literal, since there need not even be a region that the line is in. Further,
Figure 8.10: Motivating IDA
all lines may be in regions, but if a line does not depict an edge, then there can also be no positive $\Lambda$ literals.

The following examples illustrate the relationship between $IDA$ and the extension of $\Lambda$:

**Example:** Let $I$ be the image in Figure 8.10(a). There is a model $\mathcal{M}$ of $T_{\text{cardworld}} \cup I \cup IDA$ such that

$$\{s_1, s_2\} \subseteq \mathcal{M}$$

such that

$$\langle r_1, s_1 \rangle, \langle r_2, s_2 \rangle \in \Delta$$

$$\langle l_1, r_1 \rangle, \langle l_2, r_2 \rangle, \langle l_3, r_2 \rangle, \langle l_4, r_2 \rangle.$$ 

$$\langle l_5, r_2 \rangle, \langle l_6, r_1 \rangle, \langle l_7, r_1 \rangle, \langle l_8, r_1 \rangle, \langle l_9, r_1 \rangle, \langle l_{10}, r_1 \rangle \in \Lambda$$

In this case, all depicted edges correspond to lines in the extension of the $\Lambda$ relation. □

**Example:** Let $I$ be the image in Figure 8.10(b). There is a model $\mathcal{M}$ of $T_{\text{cardworld}} \cup I \cup IDA$ with

$$\{s_1, s_2\} \subseteq \mathcal{M}$$

such that

$$\langle r_1, s_1 \rangle, \langle r_1, s_2 \rangle \in \Delta$$

$$\langle l_1, r_1 \rangle, \langle l_2, r_1 \rangle, \langle l_3, r_1 \rangle, \langle l_4, r_1 \rangle,$$

$$\langle l_5, r_1 \rangle, \langle l_6, r_1 \rangle, \langle l_7, r_1 \rangle \in \Lambda$$

Note that this image violates $URDA$: the same scene depicted with no lacunary errors in edge detection is Figure 8.10a.

Since $\mathcal{M}$ satisfies $IDA$, this example also shows that $URDA$ is independent of $IDA$. □

Now consider the image in Figure 8.10(c); there are no regions in this image, and consequently there are no depicted surfaces. For this image, we cannot define any positive $\Lambda$ literals corresponding to each edge.

**Example:** Let $I$ be the image in Figure 8.10(d). There is a model $\mathcal{M}$ of $T_{\text{cardworld}} \cup I$ with

$$\{e_1, e_2, s_1\} \subseteq \mathcal{M}$$

such that

$$\langle r_1, s_1 \rangle \in \Delta$$

$$\langle l_1, r_1 \rangle, \langle l_2, r_1 \rangle, \langle l_3, r_1 \rangle, \langle l_4, r_1 \rangle \in \Lambda$$

$$\langle l_2, e_2 \rangle \in \Delta, \langle e_2, s_1 \rangle \in \text{part}$$

In this case, we have an edge denoted by $e_2$ that is depicted, yet the surface containing the edge is not depicted. Thus, $IDA$ is not satisfied by $\mathcal{M}$. Figure 8.10e shows the same scene but depicted without any errors in edge detection. □
8.4.2 Lines and Region Existence

We now need to characterize the conditions necessary to guarantee IDA. First notice that IDA actually entails several different independent assumptions which we now make explicit.

IDA can be rewritten as

\[(\forall l, e, s) \text{ part}(e, s) \land \Delta(l, e) \supset (\exists r, r') \Delta(r, s) \land \text{in}(l, r') \land r = r'\]

which entails

\[(\forall l, e, s) \text{ part}(e, s) \land \Delta(l, e) \supset (\exists r) \text{in}(l, r)\]

and

\[(\forall l, e, s) \text{ part}(e, s) \land \Delta(l, e) \supset (\exists r) \Delta(r, s)\]

We will give these new assumptions names as follows:

\textbf{Definition 8.13} The Region Existence Assumption (REA) is the sentence

\[(\forall l) \text{line}(l) \supset (\exists r) \text{in}(l, r)\]

i.e., every line in an image is contained in a region.

This is an image assumption, since it applies to all lines and regions in an image. It is not in itself related to depiction. If we assume NSEA we can restrict REA to lines that depict edges, as in the following assumption.

\textbf{Definition 8.14} The Weak Region Existence Assumption (WREA) is the sentence

\[(\forall l, e) \text{line}(l) \land \Delta(l, e) \supset (\exists r) \text{in}(l, r)\]

i.e., every line depicting an edge is contained in some region.

Thus, in models satisfying WREA, there may exist lines which are not contained in regions, but such lines must be spurious noise (i.e., they cannot depict edges). It is easy to see that WREA is weaker than REA, but that if we assume that there are no spurious errors in edge detection, then these two assumptions are equivalent. The following proposition is obvious from the definitions:

\textbf{Proposition 8.4}

\[T_{\text{cardworld}} \models \text{REA} \supset \text{WREA}\]

\[T_{\text{cardworld}} \models \text{NSEA} \supset \text{REA} \equiv \text{WREA}\]

The other sentence entailed by IDA will be referred to as

\textbf{Definition 8.15} The Depicted Kernel Region Existence (DKRE) assumption is the sentence

\[(\forall l, e, s) \text{part}(e, s) \land \Delta(l, e) \supset (\exists r) \Delta(r, s)\]

i.e., for every line depicting an edge in some surface, there exists a region depicting the surface.

The fact that these two assumptions are weaker than IDA follows easily from their definitions:
Theorem 8.7

\[ T_{\text{cardworld}} \models IDA \supset WREA \]

Theorem 8.8

\[ T_{\text{cardworld}} \models IDA \supset DKRE \]

There is also an interesting relationship between \( DKRE \) and \( DKA \), that we can derive easily from the following definition.

**Definition 8.16** The Strong Depicted Kernel Assumption (SDKA) is the sentence

\[ (\forall s) \text{surface}(s) \supset (\exists r) \Delta(r, s) \]

*i.e., every surface is depicted by some region.*

**Proposition 8.5**

\[ T_{\text{cardworld}} \models DKRE \land DKA \supset SDKA \]

Note that \( DKRE \) alone does not entail \( SDKA \), since without \( DKA \) there is no guarantee that any edges in the surface are depicted. Also note that the above assumptions are independent of each other, as we can see from the following examples.

**Example:** Let \( I \) be the image in Figure 8.10(c). There is a model \( \mathcal{M} \) of \( T_{\text{cardworld}} \cup \mathcal{I} \) with \( \{e_1, s_1\} \subset \mathcal{M} \) that satisfies \( DKA \land \neg DKRE \) such that

\[ \langle s_1 \rangle \in \text{surface}, \langle e_1, s_1 \rangle \in \text{part}, \langle l_1, e_1 \rangle \in \Delta \]

yet there does not exist a region which depicts \( s_1 \). \( \Box \)

**Example:** Let \( I \) be the image in Figure 8.10(f). There is a model \( \mathcal{M} \) of \( T_{\text{cardworld}} \cup I \) such that \( \{e_1, e_2, s_1, s_2\} \subset \mathcal{M} \) and that satisfies \( WREA \land \neg DKRE \) such that

\[ \langle s_1 \rangle, \langle s_2 \rangle \in \text{surface}, \langle e_1, s_1 \rangle, \langle e_2, s_2 \rangle \in \text{part} \]

\[ \langle l_2, e_1 \rangle, \langle l_2, e_2 \rangle, \langle r_1, s_1 \rangle \in \Delta \]

yet there does not exist a region that depicts \( s_2 \). Note that the line \( l_1 \) does not depict an edge. \( \Box \)

**Example:** Let \( I \) be the image in Figure 8.10(g), where

\[ I \models \neg(\exists r) \text{in}(l_1, r) \]

There is a model \( \mathcal{M} \) of \( T_{\text{cardworld}} \cup \mathcal{I} \) such that \( \{e_1, s_1, s_2\} \subset \mathcal{M} \) and that satisfies \( DKRE \land \neg WREA \) such that

\[ \langle r_1, s_1 \rangle, \langle r_2, s_2 \rangle, \langle l_1, e_1 \rangle \in \Delta \]

\( \Box \)
CHAPTER 8. REASONING ABOUT OCCLUSION WITH NO ERRORS IN EDGE DETECTION

Figure 8.11: Examples of IDA
It is important to observe that the converses of the above propositions do not hold. If we recall the form of $IDA$, there were three conjuncts — the first conjectured the existence of a region depicting the surface ($DKRE$), the second conjectured the existence of a region containing the line ($WRE$), and the third conjectured the equality of these two regions.

**Example:** Let $I$ be the image in Figure 8.11(a). There is a model $\cal M$ of $\text{cardworld} \cup I$ such that

\[
\{e_1, s_1, s_2\} \subseteq M \quad \text{and such that}
\]

\[
\langle r_1, s_1 \rangle, \langle r_2, s_2 \rangle, \langle l_1, e_1 \rangle, \langle l_1, e_1 \rangle \in \Delta
\]

\[
\langle e_1, s_2 \rangle \in \text{part}
\]

\[
\langle e_1, s_1 \rangle \in \text{occluding}
\]

It is easy to see that $\cal M \models WREA \land SDKA \land \neg IDA$, since $l_1$ depicts an edge in $s_2$, yet it is not in a region depicting $s_2$. This scene is illustrated in (b) with all undepicted edges displayed. 

**Example:** Let $I$ be the image in Figure 8.11(c). There is a model $\cal M$ of $\text{cardworld} \cup I$ such that

\[
\{e_1, s_1, s_2\} \subseteq M \quad \text{and such that}
\]

\[
\langle r_1, s_1 \rangle, \langle r_2, s_2 \rangle, \langle l_1, e_1 \rangle, \langle l_1, e_2 \rangle \in \Delta
\]

\[
\langle e_1, s_2 \rangle, \langle e_2, s_2 \rangle \in \text{part}
\]

\[
\langle e_1, e_2, t_2 \rangle \in \text{abuts}
\]

Note that $\cal M \models WREA \land DKRE \land \neg IDA$, since $l_1$ depicts an edge in $s_2$, yet it is not in a region depicting $s_2$. This scene is illustrated in (d) with all undepicted edges displayed.

**Example:** Another example is the image $I$ in Figure 8.12(a) where there is a model $\cal M$ of $\text{cardworld} \cup I$ such that $\{e_1, e_2, e_3, s_1, s_2, s_3, s_4\} \subseteq M$ and that satisfies $WREA \land DKRE \land \neg IDA$, but which has no abutting or occluding surfaces:

\[
\langle r_2, s_1 \rangle, \langle r_3, s_2 \rangle, \langle r_4, s_3 \rangle, \langle r_5, s_4 \rangle \in \Delta
\]

\[
\langle l_1, e_1 \rangle, \langle l_2, e_2 \rangle, \langle l_3, e_3 \rangle, \langle l_4, e_4 \rangle \in \Delta
\]

\[
\langle e_1, s_1 \rangle, \langle e_2, s_2 \rangle, \langle e_3, s_3 \rangle, \langle e_4, s_4 \rangle \in \text{part}
\]

In this model, the region $r_1$ does not depict a surface at all. This scene is illustrated in Figure 8.12(b) with all undepicted edges displayed. 

Figure 8.12: More examples of IDA
8.4.3 The Multiple Background Region Existence Assumption

In all of the examples of the preceding section, the region containing the line depicting the edge exists, and the region depicting the surface that contains the edge exists, but these regions are distinct. Thus simply asserting the existence of regions is insufficient. The important aspect of IDA is that in conjecturing the existence of regions, it also assigns edge and surface depiction literals so that the edges depicted by the lines in the image are grouped with the correct surfaces. There is thus an additional assumption implicit in IDA.

The problem is that even though the line depicting the edge is in a region, that region does not depict the surface containing the edge; furthermore, the line is only in that region, so that IDA is falsified. The assumption we need must conjecture the existence of an additional distinct region in this case which will satisfy IDA.

Definition 8.17 The Multiple Background Region Existence assumption (MBRE) is the sentence

\[(\forall l, r, e, s) \text{in}(l, r) \land \Delta(l, e) \land \text{part}(e, s) \land \neg \Delta(r, s) \supset (\exists r') \text{in}(l, r') \land r \neq r'\]

i.e., if a line that depicts an edge in some surface is contained in a region that does not depict that surface, then there must exist another region that contains the line.

This assumption guarantees that if a line depicts an edge, but is contained in a region that does not depict the surface containing the edge, then the line is contained in two distinct regions.

We can easily show that this assumption is, in fact, entailed by IDA:

Theorem 8.9

\[T_{\text{cardworld}} \models \text{IDA} \supset \text{MBRE}\]

Proof: Suppose MBRE is falsified:

\[T_{\text{cardworld}} \cup \neg \text{MBRE} \models (\exists l, r, e, s) \text{in}(l, r) \land \Delta(l, e) \land \text{part}(e, s) \land \neg \Delta(r, s) \land ((\forall r') \text{in}(l, r') \supset r = r')\]

which entails

\[T_{\text{cardworld}} \cup \neg \text{MBRE} \models (\exists l, e, s) \Delta(l, e) \land \text{part}(e, s) \land \neg (\exists r) \Delta(r, s) \land \text{in}(l, r)\]

which falsifies IDA. \[\square\]

We can now show that we have a complete characterization of IDA:

Theorem 8.10

\[T_{\text{cardworld}} \models \text{WREA} \land \text{MBRE} \land \text{DKRE} \supset \text{IDA}\]

Proof: DKRE entails

\[T_{\text{cardworld}} \cup \text{DKRE} \cup \text{WREA} \cup \text{MBRE} \models (\forall l, e) \Delta(l, e) \land \text{part}(e, s) \supset (\exists r) \Delta(r, s)\]

By WREA we have

\[T_{\text{cardworld}} \cup \text{DKRE} \cup \text{WREA} \cup \text{MBRE} \models (\forall l, e) \Delta(l, e) \supset (\exists r) \text{in}(l, r)\]
If the regions in these two sentences are the same, then together they are equivalent to IDA. Now suppose that the regions in the previous two sentences are distinct:

$$T_{\text{cardworld}} \cup DKRE \cup WREA \cup MBRE \models (\forall l, e) \Delta(l, e) \land \text{part}(e, s) \supset (\exists r, r') \Delta(r, s) \land \neg \Delta(r', s) \land \text{in}(l, r') \land r \neq r'$$

By MBRE we get

$$T_{\text{cardworld}} \cup DKRE \cup WREA \cup MBRE \models (\forall l, e) \Delta(l, e) \land \text{part}(e, s) \supset (\exists r', r'') \text{in}(l, r') \land \text{in}(l, r'') \land r' \neq r''$$

By the depiction axioms in $T_{\text{kernel}}$, one of the regions must depict the surface:

$$T_{\text{cardworld}} \cup DKRE \cup WREA \cup MBRE \models \Delta(r', s) \lor \Delta(r'', s)$$

Thus, one of the regions containing the line must depict the surface containing the edge:

$$T_{\text{cardworld}} \cup DKRE \cup WREA \cup MBRE \models \Delta(r', s) \land \text{in}(l, r') \lor \Delta(r'', s) \land \text{in}(l, r'')$$

and IDA is satisfied. □

### 8.5 Lacunate Errors and Region Existence

All of the assumptions we have considered to this point conjecture the existence of regions. Are there other assumptions that conjecture region existence but which are independent of IDA? In this section we will show that IDA is indeed entailed by NLEA; in the process, we will need to introduce new depiction assumptions related to region existence (this is Phase 2 in the characterization of region existence assumptions in Figure 8.8).

#### 8.5.1 Bounded Region Existence Assumption

If we consider the depiction assumptions WREA and MBRE we notice that they guarantee the existence of regions depicting occluding and abutting surfaces. However, as the following example indicates, there may models consistent with IDA which still violate NLEA because there are positions interior to a surface that are not included in any region:

Example: Let $I$ be the image in Figure 8.13(a). There is a model $M$ of $T_{\text{cardworld}} \cup I$ such that

$$\{e_1, e_3, s_1, s_2\} \subset M$$

and such that

$$\langle r_1, s_1 \rangle, \langle r_2, s_2 \rangle, \langle l_1, e_1 \rangle \in \Delta$$

$$\langle e_1, s_1 \rangle, \langle e_3, s_2 \rangle \in \text{part}$$

$$\langle s_1, t_1 \rangle \in \text{interior}$$

but there does not exist a line that depicts edge $e_3$. Thus $M \models IDA \land \neg NLEA$, since there are edges in $s_1$ that are undepicted and unoccluded. The image in Figure 8.13(d) depicts the same scene, but without lacunate errors. □
Figure 8.13: Examples of BREA
The problem with the model for the image in Figure 8.13(a) is that $s_2$ is not an occluding surface, but rather it is an occluded surface; \textit{IDA} is therefore not applicable to conjecture the existence of the region depicting the unoccluded parts of the surface $s_2$. To eliminate examples such as this, we will therefore need a stronger assumption.

Several ideas motivate the required assumption. Since regions are depicted surfaces, it should define how properties of surfaces are depicted in images. In the previous section, the depiction assumption \textit{INA} conjectured that the depiction relation for \textit{interior} and \textit{inclusion} was one of equivalence, given that the region exists. A reasonable approach, then, would be an assumption that guarantees that such a region exists, expressed in terms of the predicates \textit{interior} and \textit{inclusion}.

\textbf{Definition 8.18} \textit{The Bounded Region Existence Assumption (BREA) is the sentence}

$$(\forall s, t) \text{interior}(s, t) \supset (\exists r) \text{inclusion}(r, t) \vee (\exists q, p, l) \text{part}(p, s) \wedge \Delta(q, p) \wedge \text{border}(l, r) \wedge \text{position}(p, t)$$

\textit{i.e., every position interior to a surface is either included in a region or it is the position of a pixel depicting a point in the surface, and this position is in the border of the region.}

Notice that this would entail the existence of a region in Figure 8.13(a), since all positions interior to a surface must be included in some region.

It is straightforward to show that this new assumption is stronger than \textit{IDA}:

\textbf{Theorem 8.11} \textit{For any planar image $I$,}

$$T_{\text{cardworld}} \cup I \models \text{BREA} \supset \text{IDA}$$

\textbf{Proof:} We will prove this by showing that \textit{BREA} entails \textit{WREA.SDKA}, and \textit{MBREA}, which we have shown to be equivalent to \textit{IDA}.

We first consider \textit{WREA}:

$$T_{\text{cardworld}} \cup I \cup \neg \text{WREA} \models (\exists l, e) \text{line}(l) \wedge \Delta(l, e) \wedge \neg (\exists r) \text{in}(l, r)$$

Because $I$ is planar, we get

$$T_{\text{cardworld}} \cup I \cup \neg \text{WREA} \models (\exists l, e) \text{line}(l) \wedge \Delta(l, e) \wedge \neg (\exists r, t) (\text{in}(l, r) \wedge \text{border}(l, r) \wedge \text{inclusion}(r, t))$$

so that \textit{BREA} is falsified.

Next, we consider \textit{MBREA}:

Suppose \textit{MBREA} is false:

$$T_{\text{cardworld}} \cup I \cup \neg \text{BREA} \models (\exists l, r, e, s) \text{in}(l, r) \wedge \Delta(l, e) \wedge \text{part}(e, s) \wedge \neg \Delta(r, s) \wedge \neg (\exists r') \text{in}(l', r') \wedge r \neq r'$$

By $T_{\text{interior}}$, we have

$$T_{\text{cardworld}} \cup I \cup \neg \text{BREA} \models (\exists l, r, e, s) \text{in}(l, r) \wedge \Delta(l, e) \wedge \text{part}(e, s) \wedge \neg \Delta(r, s) \wedge \neg (\exists r') \text{in}(l', r') \wedge r \neq r'$$

$$\wedge (\exists t) \text{boundary}(e, t) \wedge \text{interior}(s, t)$$

By \textit{BREA} and axiom 6.13, we have

$$T_{\text{cardworld}} \cup I \cup \neg \text{BREA} \models (\exists l, r, e, s) \text{in}(l, r) \wedge \Delta(l, e) \wedge \text{part}(e, s) \wedge \neg \Delta(r, s) \wedge \neg (\exists r') \text{in}(l', r') \wedge r \neq r'$$

$$\wedge (\exists t) \text{boundary}(e, t) \wedge \text{interior}(s, t) \wedge (\exists r_2) \text{inclusion}(r_2, t) \wedge \text{border}(l, r_2)$$
Since the edge is depicted, by axiom 7.16, there exists a position at which the surface cannot be obscured by another surface:

\[ T_{\text{cardworld}} \cup I \cup \text{BREA} \models \exists l, r, e, s . in(l, r) \land \Delta(l, e) \land \text{part}(e, s) \land -\Delta(r, s) \land -(\exists r') \text{in}(l, r') \land r \neq r' \]
\[ \land (\exists t) \text{boundary}(e, t) \land \text{interior}(s, t) \land (\exists r_2) \text{inclusion}(r_2, t) \land \text{border}(l, r_2) \land -(\exists s') \text{obsures}(s', s, t) \]

By axiom 7.17, this new region must depict the surface:

\[ T_{\text{cardworld}} \cup I \cup \text{BREA} \models \exists l, r, e, s . in(l, r) \land \Delta(l, e) \land \text{part}(e, s) \land -\Delta(r, s) \land -(\exists r') \text{in}(l, r') \land r \neq r' \]
\[ \land (\exists t) \text{boundary}(e, t) \land \text{interior}(s, t) \land (\exists r_2) \text{inclusion}(r_2, t) \land \text{border}(l, r_2) \land \Delta(r_2, s) \]

But this contradicts our assumption that \( \text{MBRE} \) is false, since the regions \( r \) and \( r_2 \) are distinct (one region depicts the surface, and the other does not).

It is trivial to see that \( \text{BREA} \) entails \( \text{SDKA} \). \( \square \)

In order to prove that \( \text{BREA} \) is entailed by \( \text{NLEA} \), we need to show that \( \text{NLEA} \) guarantees the existence of regions, even though it only refers to the depiction and occlusion of points. To accomplish this, we will use the properties of the planar images presented in Chapter 6.2, which specified the conditions for region existence within an image, and which also defined the extension of \text{inclusion} for regions.

We will first need the following definition.

**Definition 8.19** An edge is a top edge for a surface iff

\[(\forall e, s) \text{top}(e, s) \equiv \]

\[ ((\exists p, t) \text{part}(e, s) \land \text{part}(p, e) \land \text{position}(p, t) \land -(\exists s^*) \text{obsures}(s^*, s, t)) \]
\[ \lor ((\exists s_1, p, t) \text{occluding}(e, s) \land \text{part}(e, s_2) \land \text{part}(p, e) \land \text{position}(p, t) \land -(\exists s^*) \text{obsures}(s^*, s_1, t)) \]
\[ \land ((\forall s') s \neq s_2 \land s \neq s' \supset (\text{obsures}(s, s', t) \equiv \text{obsures}(s_2, s'', t))) \]

Thus a top edge \( e \) for a surface \( s \) is either an unobscured edge that is part of the surface, or it is an occluding edge that is unobscured and which obscures the same set of surfaces as \( s \). This last condition is necessary to avoid situations such as Figure 8.14(b).

**Example:** Let \( I \) be the image in Figure 8.14(b). There is a model \( M \) of \( T_{\text{cardworld}} \cup \text{NLEA} \cup I \) such that \( \{e_1, s_1, s_2, s_3\} \subset M \) and that satisfies

\[ \langle r_1, s_1 \rangle, \langle r_2, s_2 \rangle, \langle r_3, s_3 \rangle, \langle l_1, e_1 \rangle \in \Delta \]
\[ \langle e_1, s_2 \rangle \in \text{part}, \langle e_1, s_1 \rangle \in \text{occluding} \]

In this model \( l_1 \) should not depict a top edge since it is also an occluding edge for an intervening surface, namely \( s_2 \). Intuitively, we want the depicted top edges to form a set of intersecting lines that are the border for a region depicting \( s_1 \); if the line depicts an edge that occludes some other intervening surface, it will not satisfy this intuition. \( \square \)
Figure 8.14: Examples for top edges.
The approach for proving the theorem can be seen by considering the image in Figure 8.14(a). Lines $l_1, l_2, l_7, l_9, l_{12}$ depict edges in the surface, while the remaining lines depict occluding edges: all of these lines therefore depict top edges. Since $NLEA$ entails $WCCA$ all of these lines must intersect each other: given this set of intersecting lines, we can entail the existence of a region using the properties of planar images.

**Theorem 8.12** For any planar image $I$,

$$T_{\text{cardworld}} \cup I \models NLEA \supset BREA$$

**Proof:** The proof for this theorem has four stages:

1. Top edges exist in any scene.
2. Top edges are always depicted.
3. The line which are depicted top edges intersect other lines which are also depicted top edges.
4. Depicted top edges are the border lines of a region.

As with other theorems, we will prove this theorem by cases depending on whether the position interior to the surface is obscured or not obscured by another surface.

**Case 1:** Consider those positions that satisfy

$$T_{\text{cardworld}} \cup I \models (\exists s, t) \text{interior}(s, t) \land \neg(\exists s') \text{obscures}(s', s, t)$$

We will first show that depicted top edges exist in any scene. We then show that by $NLEA$, top edges are always depicted and that the lines depicting top edges form a region.

**Lemma 8.3**

$$T_{\text{cardworld}} \models (\forall s, t_1) \text{interior}(s, t_1) \supset (\exists e, p, t_2) \text{top}(e, s) \land \text{position}(p, t_2) \land \text{interior}(s, t_2) \land \text{noncrossing}(s, t_1, t_2)$$

**Proof:** By the definition of interior,

$$T_{\text{cardworld}} \models (\forall s, t_1) \text{interior}(s, t_1) \supset (\exists e, p, t_2) \text{boundary}(e, t_1) \land \text{part}(p, e) \land \text{position}(p, t_2) \land \text{interior}(s, t_2) \land \text{noncrossing}(s, t_1, t_2)$$

Either this boundary edge is unobscured, or there exists an occluding edge in the occluding surface which obscures the boundary edge (by the Oculding Edge Existence Lemma). By axiom 7.10 of $T_{\text{obscures}}$, there must exist an occluding edge which is unobscured. Further, there must exist such an occluding edge which occludes only the surface $s$, since by the hypothesis of Case 1, $s$ is unobscured at the position $t_1$. Since the $\text{obscures}$ ordering is discrete, if $s$ is partially occluded, then at any position where $s$ is obscured, there must exist a surface which obscures $s$ with no intervening obscuring surface. Thus, the edge satisfies the definition of $\text{top}$. □

**Lemma 8.4** For any planar image $I$,

$$T_{\text{cardworld}} \cup I \cup NLEA \models (\forall e, s) \text{top}(e, s) \supset (\exists l) \Delta(l, e)$$
Proof: This follows from the definitions of $N LEA$ and $\text{top}$. \hfill $\square$

Lemma 8.5 For any planar image $I$,

$$T_{\text{cardworld}} \cup I \cup N LEA \models (\forall l, e, s) \top(e, s) \land \Delta(l, e) \supset (\exists l', l'', e', e'', j, j') \text{intersect}(l, l', j) \land \text{intersect}(l, l'', j') \land \Delta(l', e') \land \Delta(l'', e'') \land \top(e', s) \land \top(e'', s)$$

i.e., the lines which are depicted top edges intersect other lines which are also depicted top edges.

Proof: Since $T_{\text{cardworld}} \cup I \models N LEA \supset WCCA$ we have

$$T_{\text{cardworld}} \cup I \cup N LEA \models (\forall l, e, s) \top(e, s) \land \Delta(l, e) \supset (\exists p, t) \text{part}(p, e) \land \text{position}(p, t) \land \text{interior}(s, t)$$

Thus if a line depicting a top edge intersects a line that does not depict a top edge, then we have

$$T_{\text{cardworld}} \cup I \cup N LEA \models (\forall e, e', s, l, l', q, t) \top(e, s) \land \neg \top(e', s) \land \Delta(l, e) \land \Delta(l', e') \land \text{intersect}(l, l', q) \land \text{position}(q, t) \supset \neg \text{interior}(s, t)$$

In this case, the top edge cannot be an edge in the occluded surface: by the definition of $\text{top}$ it must be an occluding edge. Thus, there must be a point in an edge of the occluded surface that is obscured by this occluding top edge:

$$T_{\text{cardworld}} \cup I \cup N LEA \models (\forall e, e', s, l, l', q, t) \top(e, s) \land \neg \top(e', s) \land \Delta(l, e) \land \Delta(l', e') \land \text{intersect}(l, l', q) \land \text{position}(q, t) \land \neg \text{interior}(s, t)$$

$$\supset (\exists e'', p, t_2) \text{occluding}(e, s) \land \text{part}(p, e'') \land \text{part}(e'', s) \land \text{position}(p, t_2) \land (\exists t_3) \neg (\exists s') \text{obscures}(s', s, t_3)$$

By $N LEA$ this edge must be depicted and by $WCCA$ it must intersect the line depicting the occluding edge. Since both $e$ and $e'$ are top edges, we have proven the lemma. \hfill $\square$

Thus all lines depicting top edges for a surface intersect lines depicting other top edges for that surface. The next lemma shows that this set of intersecting lines forms a region.

Lemma 8.6 For any planar image $I$,

$$T_{\text{cardworld}} \cup I \cup N LEA \models (\forall l, e, s) \top(e, s) \land \Delta(l, e) \supset (\exists r) \text{in}(l, r) \land \text{border}(l, r)$$

Proof: By induction on the set of top edges:

Base case: $n = 3$. By the preceding lemma, we have

$$T_{\text{cardworld}} \cup I \cup N LEA \models (\exists e_1, e_2, e_3, s, l_1, l_2, j_1, j_2, j_3) \top(e_1, s) \land \top(e_2, s) \land \top(e_3, s) \land \Delta(l_1, e_1) \land \Delta(l_2, e_2) \land \Delta(l_3, e_3) \land \text{intersect}(l_1, l_2, j_1) \land \text{intersect}(l_2, l_3, j_2) \land \text{intersect}(l_3, l_1, j_3)$$

Within a planar image, any cycle of intersecting lines defines a region (recall the definition of planar images in Chapter 6). Since $I$ is planar, we have
Induction step:
Suppose that we have a set of $n$ top edges:

$$T_{\text{cardworld}} \cup I \cup \text{NLEA} \models (\forall e_1, e_2, \ldots, e_n, s, l_1, l_2, \ldots, l_n, j_1, \ldots, j_n) \text{top}(e_1, s) \land \cdots \land \text{top}(e_n, s)$$
$$\land \Delta(l_1, e_1) \land \Delta(l_2, e_2) \land \Delta(l_n, e_n) \land \text{intersect}(l_1, l_2, j_1) \land \cdots \land \text{intersect}(l_{n-1}, l_n, j_{n-1})$$
$$\lor (\exists r) \text{in}(l_1, r) \land \cdots \land \text{in}(l_n, r) \land \text{in}(l_n, r)$$

By the preceding lemma, we get

$$T_{\text{cardworld}} \cup I \cup \text{NLEA} \models (\forall e_1, \ldots, e_n, s, l_1, l_2, \ldots, l_n, j_1, \ldots, j_n) \text{top}(e_1, s) \land \cdots \land \text{top}(e_n, s)$$
$$\land \Delta(l_1, e_1) \land \cdots \land \Delta(l_n, e_n) \land \text{intersect}(l_1, l_2, j_1) \land \cdots \land \text{intersect}(l_{n-1}, l_n, j_{n-1}) \land \text{intersect}(l_n, l_1, j_n)$$

so that we have a cycle of intersecting lines.

Within a planar image, any cycle of intersecting lines defines a region: since $I$ is planar, we have

$$T_{\text{cardworld}} \cup I \cup \text{NLEA} \models (\forall e_1, \ldots, e_n, s, l_1, l_2, \ldots, l_n, j_1, \ldots, j_n) \text{top}(e_1, s) \land \cdots \land \text{top}(e_n, s)$$
$$\land \Delta(l_1, e_1) \land \cdots \land \Delta(l_n, e_n) \land \text{intersect}(l_1, l_2, j_1) \land \cdots \land \text{intersect}(l_{n-1}, l_n, j_{n-1}) \land \text{intersect}(l_n, l_1, j_n)$$
$$\lor (\exists r) \text{in}(l_1, r) \land \cdots \land \text{in}(l_n, r)$$

By the definition of inclusion and the previous lemmas we get

$$T_{\text{cardworld}} \cup I \cup \text{NLEA} \models (\forall s, s', t) \text{interior}(s', t) \land \neg (\exists s) \text{obsures}(s, s', t)$$
$$\lor (\exists r) \text{inclusion}(r, t) \lor (\exists q \land \text{border}(q, r) \land \text{position}(q, t))$$

At this stage, we have shown that any unobscured position interior to a surface is included in region. We now need to show that this property holds for obscured positions as well; this is the other case in the proof:

**Case 2:** Consider those positions that satisfy

$$T_{\text{cardworld}} \cup I \cup \text{NLEA} \models (\exists s, s', t) \text{interior}(s, t) \land \text{obsures}(s', s, t)$$

In Case 1, we showed that any unobscured position interior to a surface is included in a region. In Case 2, we need to show that this property holds for obscured positions as well.

The definition of obsures gives

$$T_{\text{cardworld}} \cup I \cup \text{NLEA} \models (\forall s, s', t) \text{interior}(s, t) \land \text{obsures}(s', s, t) \lor \text{interior}(s', t)$$

However, we have just shown that all surfaces are depicted by regions, so that the position of the point must be included in a region.

$$T_{\text{cardworld}} \cup I \cup \text{NLEA} \models (\forall s, s', t) \text{interior}(s, t) \land \text{obsures}(s', s, t) \land \text{interior}(s', t)$$
$$\lor (\exists r) (\Delta(r, s) \lor \Delta(r, s')) \lor (\exists q \land \text{border}(q, r) \land \text{position}(q, t))$$

Thus, the region including the position $t$ either depicts the original surface or it depicts the occluding surface. □

**Corollary 8.3** For any planar image $I$,

$$T_{\text{cardworld}} \cup I \models \text{NLEA} \supset \text{IDA}$$
8.5.2 Occluding Surface Depiction Assumption

Recall the original motivation for \textit{BREA}: these were the images in Figure 8.13 that satisfied \textit{IDA} \land \lnot \textit{NLEA}. Notice that they also violate \textit{BREA}. In each of these cases, there is a position interior to a surface, but which is neither included in a region, nor obscured by another surface. Some point in the boundary edges for this position is not depicted due to lacunate errors. In order to completely characterize \textit{BREA} we will therefore need one more region existence assumption. However, this assumption cannot simply conjecture the existence of regions, since in the examples of Figure 8.13 all surfaces were depicted and all lines were in regions. What we must do is assert the existence of regions for surfaces that are partially occluded, so that all parts of the surface will be included in some region.

\textbf{Definition 8.20} The \textit{Occluded Surface Depiction Assumption (OSDA)} is the sentence

\[ (\forall e, s_1, s_2, l) \text{occluding}(e, s_1) \land \text{part}(e, s_2) \land \Delta(l, e) \supset (\exists r) \text{in}(l, r) \land \left[ \Delta(r, s_1) \lor (\exists s') (s_2 \neq s') \land \Delta(r', s') \land \text{ontop}(s', s_2) \right] \]

\[ \text{i.e., every depicted occluding edge is in a region that depicts either the occluded surface or some other distinct occluding surface.} \]

Note that this assumption must also cover the possibility that the remainder of the surface is occluded so that the surface is depicted by only one region. In such a case, the line depicting the occluding edge will be in a region depicting an occluding surface.

Consider the images in Figure 8.13(c) and (e), which violate \textit{OSDA}.

\textbf{Example:} Let \( I \) be the image in Figure 8.13(c) where

\[ I \models \lnot \text{inclusion}(r_1, t_1) \land \text{in}(l_1, r_2) \]

There is a model \( \mathcal{M} \) of \( T_{\text{cardworld}} \cup I \) such that \( \{ e_1, s_1, s_2 \} \subset M \) and that satisfies

\[ (r_1, s_1), (r_2, s_2), (l_1, e_1) \in \Delta, (e_1, s_2) \in \text{part} \]

\[ (s_1, t_1) \in \text{interior}, (e_1, s_1) \in \text{occluding} \]

This model satisfies \textit{IDA}, but violates \textit{BREA} and \textit{OSDA}, since there is a position which is interior to the surface, but which is not included in any region. In this case, the occluded surface is not depicted by a region containing the line that depicts the occluding edge. \( \Box \)

\textbf{Example:} Let \( I \) be the image in Figure 8.13(e), where

\[ I \models \lnot \text{inclusion}(r_1, t_1) \land \text{in}(l_1, r_2) \land \text{in}(l_1, r_3) \land \text{in}(l_2, r_3) \]

There is a model \( \mathcal{M} \) of \( T_{\text{cardworld}} \cup I \) such that \( \{ e_1, e_2, s_1, s_2, s_3 \} \subset M \) and that satisfies

\[ (r_1, s_1), (r_2, s_2), (r_3, s_3), (l_2, e_2) \in \Delta, (e_2, s_3) \in \text{part} \]

\[ (s_1, t_1) \in \text{interior}, (e_2, s_1) \in \text{occluding} \]

Again, this model satisfies \textit{IDA}, but violates \textit{BREA} and \textit{OSDA}. \( \Box \)
The next two examples have similar images, but they satisfy both $OSDA$ and $BREA$.

**Example:** Let $I$ be the image in Figure 8.13(d), where

$$I \models inclusion(r_3, t_1) \land in(l_1, r_2) \land in(l_1, r_3)$$

There is a model $M$ of $T_{cardworld} \cup I$ such that $\{e_1, s_1, s_2\} \subset M$ and that satisfies $OSDA$:

$$\langle r_1, s_1 \rangle, \langle r_2, s_2 \rangle, \langle r_3, s_1 \rangle, \langle l_1, e_1 \rangle, \langle l_1, e_1 \rangle \in \Delta, \langle e_1, s_2 \rangle \in part$$

$$\langle e_1, s_1 \rangle \in occluding, \langle s_1, t_1 \rangle \in interior$$

In this case, there is a region that depicts the occluded surface and which contains the line that depicts the occluding edge. It is interesting to compare this image with the image in Figure 8.13(d), whose models violate $OSDA$. □

**Example:** Let $I$ be the image in Figure 8.13(f), where

$$I \models inclusion(r_3, t_4) \land inclusion(r_4, t_1)$$

$$\land in(l_1, r_2) \land in(l_1, r_3) \land in(l_2, r_3) \land in(l_2, r_4)$$

There is a model $M$ of $T_{cardworld} \cup I$ such that $\{e_1, e_2, s_1, s_2, s_3\} \subset M$ and that satisfies $OSDA$:

$$\langle r_1, s_1 \rangle, \langle r_2, s_2 \rangle, \langle r_3, s_3 \rangle, \langle r_4, s_1 \rangle, \langle l_1, e_1 \rangle, \langle l_2, e_2 \rangle \in \Delta$$

$$\langle e_1, s_2 \rangle, \langle e_2, s_3 \rangle \in part$$

$$\langle e_1, s_1 \rangle, \langle e_2, s_1 \rangle \in occluding$$

$$\langle s_1, t_4 \rangle, \langle s_1, t_1 \rangle \in interior$$

In this case, $r_3$ depicts a surface on top of $s_1$ (satisfying the first disjunct in the definition of $OSDA$) while $r_4$ depicts $s_1$ (satisfying the second disjunct in the definition of $OSDA$). □

We can see that $BREA$ should intuitively entail $OSDA$: since $BREA$ entails the existence of a region including positions interior to the occluded surface, and points within an occluding edge also have positions interior to the occluded surface, the depicted occluding edge should also be contained in such a region.

**Theorem 8.13**

$$T_{cardworld} \models BREA \supset OSDA$$

**Proof:** By the definitions of $occluding$ and $interior$ we have

$$T_{cardworld} \models (\forall e, s, p, l, t) occluding(e, s) \land \Delta(l, e) \land part(p, e) \land position(p, t) \supset interior(s, t)$$

so that by $BREA$ there exists a region including the position $t$:

$$T_{cardworld} \cup BREA \models (\forall e, s, p, l, t) occluding(e, s) \land \Delta(l, e) \land part(p, e) \land position(p, t) \land interior(s, t)$$


\( (\exists r) \text{ inclusion}(r, t) \)

By axiom 6.12 of \( T_{\text{interior}}^A \), if the region \( r \) does not depict the surface \( s \), then it must depict some other surface \( s' \) such that the position \( t \) is also interior to \( s' \):

\[
T_{\text{cardworld}} \cup \text{BREA} \models (\forall e, s, p, l, t) \text{ooccluding}(e, s) \land \Delta(l, e) \land \text{part}(p, e) \land \text{position}(p, t) \land \text{interior}(s, t) \\
\land \text{inclusion}(r, t) \land \neg \Delta(r, s) \supset \\
(\exists s') \Delta(r, s') \land \text{interior}(s', t)
\]

By axiom 7.1 of \( T_{\text{occ}} \), the surface \( s' \) obscures the surface \( s \) at the position \( t \):

\[
T_{\text{cardworld}} \cup \text{BREA} \models (\forall e, s, p, l, t) \text{ooccluding}(e, s) \land \Delta(l, e) \land \text{part}(p, e) \land \text{position}(p, t) \land \text{interior}(s, t) \\
\land \text{inclusion}(r, t) \land \neg \Delta(r, s) \land \Delta(r, s') \land \text{interior}(s', t) \\
\supset \text{obscures}(s', s, t) \land \Delta(r, s') \land \text{ontop}(s', s)
\]

\( \square \)

To prove the converse direction (i.e. that OSDA and IDA together entail BREA), we will need the following lemma:

**Lemma 8.7**

\[
T_{\text{cardworld}} \cup \text{NSEA} \models (\forall r, s, t) \Delta(r, s) \land \text{interior}(s, t) \land \text{exclusion}(r, t) \supset \\
(\exists l, p, e, t_2) \text{in}(l, r) \land \Delta(l, e) \land \text{part}(p, e) \land \text{position}(p, t_2) \land \text{ooccluding}(e, s)
\]

i.e., assuming that there are no spurious errors in edge detection, if a position interior to a surface is not included in a region depicting the surface, then there exists a line in the region that depicts an occluding edge.

**Proof:** By axiom 6.13 of \( T_{\text{interior}}^A \)

\[
T_{\text{cardworld}} \cup \text{NSEA} \models (\forall r, s, t) \Delta(r, s) \land \text{interior}(s, t) \land \text{exclusion}(r, t) \supset \\
(\exists l) \text{border}(l, r) \land (\neg (\exists e) \Delta(l, e) \land \text{part}(e, s))
\]

By NSEA, all border lines of the region must depict edges in some surface, so we have

\[
T_{\text{cardworld}} \cup \text{NSEA} \models (\forall r, s, t) \Delta(r, s) \land \text{interior}(s, t) \land \text{exclusion}(r, t) \supset \\
(\exists l, e, s_2) \text{border}(l, r) \land \Delta(l, e) \land \neg \text{part}(e, s) \land \text{part}(e, s_2)
\]

By the definition of occluding edges, we have

\[
T_{\text{cardworld}} \cup \text{NSEA} \models (\forall r, s, t) \Delta(r, s) \land \text{interior}(s, t) \land \text{exclusion}(r, t) \supset \\
(\exists l, p, e, t_2) \text{in}(l, r) \land \Delta(l, e) \land \text{part}(p, e) \land \text{position}(p, t_2) \land \text{ooccluding}(e, s)
\]

\( \square \)

Notice that we need the assumption NSEA to prove this lemma. Without it, spurious lines within the image could be border lines of a region without requiring the existence of occluding edges to be depicted by the border lines.

We can now complete phase 2 in characterizing the region existence assumptions, by showing that BREA is equivalent to the conjunction of IDA and OSDA:
Theorem 8.14

\[ T_{\text{cardworld}} \cup NSE = IDA \land OSDA \supset BREA \]

Proof: We will show \( \neg BREA \land IDA \supset \neg OSDA \).

By the definition of \( BREA \) we have

\[ T_{\text{cardworld}} \cup \neg BREA \land IDA \models (\exists s, t) \text{interior}(s, t) \land \neg (\exists r) \text{inclusion}(r, t) \]

Lemma 8.7 then gives

\[ T_{\text{cardworld}} \cup \neg BREA \land IDA \models (\exists s, t) \text{interior}(s, t) \land \neg (\exists r) \text{inclusion}(r, t) \land (\exists s') \text{ontop}(s', s) \]

By \( IDA \),

\[ T_{\text{cardworld}} \cup \neg BREA \land IDA \models (\exists s, t) \text{interior}(s, t) \land \neg (\exists r) \text{inclusion}(r, t) \land ((\forall s') \text{ontop}(s', s) \supset (\exists r') \Delta(r', s')) \]

The original position is not included in any of the regions depicting the occluding surfaces since \( BREA \) is falsified:

\[ T_{\text{cardworld}} \cup \neg BREA \land IDA \models (\exists s, t) \text{interior}(s, t) \land \neg (\exists r) \text{inclusion}(r, t) \land ((\forall s', r') \text{ontop}(s', s) \land \Delta(r', s') \supset \neg \text{inclusion}(r', t)) \]

By the preceding lemma, there must exist a depicted occluding edge for the original surface; by the preceding sentence the line depicting the occluding edge must be in a unique region (the region depicting the occluding surface containing the occluding edge) otherwise the position would be included in one of the regions containing the line:

\[ T_{\text{cardworld}} \cup \neg BREA \land IDA \models (\exists s, t) \text{interior}(s, t) \land \neg (\exists r) \text{inclusion}(r, t) \land ((\forall s', r') \text{ontop}(s', s) \land \Delta(r', s') \supset \neg \text{inclusion}(r', t)) \land (\exists e, l, t_2) \text{occluding}(e, s, t_2) \land \Delta(l, e) \land (\exists r'') \Delta(r'', s) \land r' \neq r'' \]

This falsifies \( OSDA \). \( \Box \)

8.5.3 Summary and Implications

In this section, we have considered assumptions which conjecture the existence of regions in an image, given that we are assuming no lacunate errors in edge detection. These assumptions about region existence are essentially specifying the necessary and sufficient conditions for the depiction of surfaces, just as \( NLEA \) specifies the necessary and sufficient conditions for the depiction of points.

As with the assumptions concerning uniquely depicting regions, the region existence assumptions show how occlusion assumptions about points are equivalent to depiction assumptions about surfaces. Intuitively, if lacunate errors in edge detection are allowed, then there could be points in edges which are not depicted, and the result would be that there is a gap in the lines depicting these edges. As we saw with the proof of \( NLEA \supset BREA \), if there were such gaps, then we would not have a set of intersecting lines, and hence there would exist no region containing the lines.
8.6 Characterizing NLEA

To this point, we have uncovered two depiction assumptions, INDA and BREA, that are entailed by NLEA: in this section, we show that these assumptions are in fact necessary and sufficient conditions for NLEA. The problem we now face is that both BREA and INDA are depiction assumptions, while NLEA is an occlusion assumption that constrains the extension of the obscures relation. If we are to fully characterize NLEA, we will need to prove that these assumptions also determine the extension of obscures.

We can first observe that URDA, the assumption about uniquely depicting regions, allows us to determine which surfaces are obscured at some position, if we know which surface is depicted by the region:

**Lemma 8.8**

\[ T_{\text{cardworld}} \cup URDA \models (\forall r, s, s', t) \Delta(r, s) \land \text{inclusion}(r, t) \land \text{interior}(s, t) \land \text{interior}(s', t) \land \neg \Delta(r, s') \supset \text{obscures}(s, s', t) \]

i.e., if a position is interior to a set of surfaces and is also included in a region that depicts a surface, then the depicted surface obscures all other surfaces at that position.

**Proof:** By axiom 7.17 of T_{occ} (the occlusion axiom for surfaces), any undepicted surface must be obscured at some position:

\[ T_{\text{cardworld}} \cup URDA \models (\forall r, s, s', t) \Delta(r, s) \land \text{inclusion}(r, t) \land \text{interior}(s, t) \land \text{interior}(s', t) \land \neg \Delta(r, s') \supset (\exists s'') \text{obscures}(s'', s', t) \]

Claim: The depicted surface is not obscured at this position:

\[ T_{\text{cardworld}} \cup URDA \models (\forall r, s, s', t) \Delta(r, s) \land \text{inclusion}(r, t) \land \text{interior}(s, t) \land \text{interior}(s', t) \land \neg \Delta(r, s') \supset \neg (\exists s) \text{obscures}(s', s, t) \]

**Proof:** Suppose

\[ T_{\text{cardworld}} \cup URDA \models (\exists r, s, s', t) \Delta(r, s) \land \text{inclusion}(r, t) \land \text{interior}(s, t) \land \text{interior}(s', t) \land \neg \Delta(r, s') \]

\[ \land (\exists s) \text{obscures}(s, s, t) \]

Then by axiom 7.10 of T_{obscures} there exists a maximal surface s* such that

\[ T_{\text{cardworld}} \cup URDA \models (\forall r, s, s', t) \Delta(r, s) \land \text{inclusion}(r, t) \land \text{interior}(s, t) \land \text{interior}(s', t) \land \neg \Delta(r, s') \supset \]

\[ \land (\exists s) \text{obscures}(s, s, t) \land (\forall s'') \neg \text{obscures}(s'', s, t) \]

and by axiom 7.17 of T_{occ} we have

\[ T_{\text{cardworld}} \cup URDA \models (\forall r, s, s', t) \Delta(r, s) \land \text{inclusion}(r, t) \land \text{interior}(s, t) \land \text{interior}(s', t) \land \neg \Delta(r, s') \supset \]

\[ (\exists r, s') \Delta(r, s') \]

contradicting URDA. □
Combining these two sentences and axiom 7.1 of $T_{occ}$ we get

$$T_{\text{cardworld}} \cup U^{RDA} \models (\forall r, s, s', t) \Delta(r, s) \land \text{inclusion}(r, t) \land \text{interior}(s, t) \land \text{interior}(s', t) \land \neg \Delta(r, s')$$

$\supset \text{obscures}(s, s', t)$

$\blacksquare$

Using this lemma to determine the extension of $\text{obscures}$ given the set of depicted surfaces, we can finally prove that the depiction assumptions $\text{BREA}$ and $\text{INDA}$ are equivalent to the occlusion assumption $\text{NLEA}$.

**Theorem 8.15** For any planar image $I$,

$$T_{\text{cardworld}} \cup I \models \text{BREA} \land \text{INDA} \supset \text{NLEA}$$

**Proof:** Suppose for some planar image $I$, that there exists an undepicted point:

$$T_{\text{cardworld}} \cup I \cup \text{BREA} \cup \text{INDA} \models (\exists p, t) \text{point}(p) \land \text{position}(p, t) \land \neg (\exists q) \Delta(q, p)$$

By the definition of $\text{interior}$, this point is part of some surface, and the point’s position is interior to this surface:

$$T_{\text{cardworld}} \cup I \cup \text{BREA} \cup \text{INDA} \models (\forall p, t) \text{point}(p) \land \neg (\exists q) \Delta(q, p) \land \text{position}(p, t)$$

$\supset (\exists s) \text{part}(p, s) \land \text{interior}(s, t)$

By $\text{BREA}$, there exists a region depicting some surface such that either the position of the point is included in the region or it is the position of a pixel in a border line of the region:

$$T_{\text{cardworld}} \cup I \cup \text{BREA} \cup \text{INDA} \models (\forall p, s, t) \text{point}(p) \land \neg (\exists q) \Delta(q, p) \land \text{position}(p, t) \land \text{part}(p, s) \land \text{interior}(s, t)$$

$\supset (\exists s') \neg \text{exclusion}(r, t) \land \Delta(r, s')$

By $\text{INDA}$, the position must also be interior to the surface depicted by the region:

$$T_{\text{cardworld}} \cup I \cup \text{BREA} \cup \text{INDA} \models (\forall p, s, t) \text{point}(p) \land \neg (\exists q) \Delta(q, p) \land \text{position}(p, t) \land \text{part}(p, s) \land \text{interior}(s, t)$$

$\supset (\exists r, s') \neg \text{exclusion}(r, t) \land \Delta(r, s') \land \text{interior}(s', t)$

Recall that the sentence in Lemma 8.2 can be written as

$$(\forall p, s, r, t) \text{part}(p, s) \land \text{position}(p, t) \land \text{inclusion}(r, t) \supset \neg \Delta(r, s)$$

so that we know the surface containing the point cannot be depicted by the region:

$$T_{\text{cardworld}} \cup I \cup \text{BREA} \cup \text{INDA} \models (\forall p, s, t) \text{point}(p) \land \neg (\exists q) \Delta(q, p) \land \text{position}(p, t) \land \text{part}(p, s) \land \text{interior}(s, t)$$

$\supset (\exists r, s') \neg \text{exclusion}(r, t) \land \Delta(r, s') \land \text{interior}(s', t) \land \neg \Delta(r, s)$

Lemma 8.8 then entails that the surface depicted by the region must be obscuring the surface containing the point:

$$T_{\text{cardworld}} \cup I \cup \text{BREA} \cup \text{INDA} \models (\forall p, s, t) \text{point}(p) \land \neg (\exists q) \Delta(q, p) \land \text{position}(p, t) \land \text{part}(p, s) \land \text{interior}(s, t)$$

$\supset (\exists r, s') \neg \text{exclusion}(r, t) \land \Delta(r, s') \land \text{interior}(s', t) \land \neg \Delta(r, s)$

$\land \text{obscures}(s', s, t)$

which entails $\text{NLEA}$. $\blacksquare$
In effect, this proof uses \textit{BREA} to guarantee the existence of regions, \textit{INDA} to restrict the properties of regions, and the occlusion axioms to assign \textit{obscures} at the positions of undepicted scene objects.

This theorem shows that the occlusion assumption \textit{NLEA} has been decomposed into a set of depiction assumptions: \textit{obscures} is then assigned using the occlusion axioms in \textit{T}_{occ}.

In Figure 8.15 we illustrate the relationship between \textit{NLEA} and the other nondepiction assumptions that have been used in this chapter. The children of a parent node in the graph are entailed by the parent, and the conjunction of children together entail the parent.

Different combinations of these assumptions can also be used for theories with errors in edge detection, each of which is a conservative extension of \textit{T}_{cardworld}. We can assume that \textit{NLEA} is not satisfied, so that one of the assumptions in Figure 8.15 must also be violated. For example, we can consider theories which allow some lacunae errors in edge detection, such as images containing silhouettes, but which still require region existence. Models of such images violate the \textit{INDA} and \textit{URDA} assumptions, but satisfy \textit{BREA}.

\section{Applications of \textit{NLEA}}

Using \textit{NLEA} allows us to strengthen many of the results we have proven about the occlusion axioms, since all undepicted scene objects must be occluded. This section will present these results.

First, the occlusion assumption \textit{NLEA} also allows us to prove a stronger version of the depicted kernel theorem. Notice that by the occlusion axioms, no part of a totally occluded surface is depicted, nor is a totally occluded surface depicted.

\textbf{Theorem 8.16} Given a planar image \(I\), any model \(M\) of \(T_{\text{cardworld}} \cup \text{NLEA} \cup I\) is an extension of a unique kernel model \(M'\) constructed by adjoining a set of totally occluded surfaces to \(M'\).
Proof: The Depicted Kernel Theorem states that we can construct any model by adjoining a set of undepicted surfaces to any kernel model. \( NLEA \) requires that all undepicted scene objects be occluded. Since the undepicted surfaces also have no depicted parts, they must be totally occluded.

\( NLEA \) also allows us to strengthen the characterization theorem for occlusion and cues for occlusion that we presented in the previous section. Consider the earlier example in Figure 8.10(b): without \( NLEA \), there was ambiguity over the \textit{obscures} ordering between \( s_1 \) and \( s_2 \). If we assume \( NLEA \), however, there is the unique assignment \( \text{obscures}(s_2, s_1, t_1) \), since \( \Delta(r_2, s_2) \) and \( NLEA \) forces \( r_2 \) to be uniquely depicting. If \( \text{obscures}(s_1, s_1, t_1) \) were consistent, then the occlusion axioms would force \( r_2 \) to depict both surfaces \( s_1 \) and \( s_2 \), violating \( URDA \).

With \( NLEA \), if there is evidence for occlusion (a depicted occluding edge), then we infer that \( s \) obscures \( s' \) at \( t \); if there is no evidence of occlusion (there is no depicted occluding edge), then we infer that \( s \) does not obscure \( s' \) at \( t \) and hence \( s' \) obscures \( s \) at \( t \). We now make these intuitions precise in the following theorem.

Theorem 8.17 For any planar image \( I \),
\[
T_{card_world} \cup I \cup NLEA \models (\forall r, s, s', t) \Delta(r, s) \land inclusion(r, t) \land interior(s', t) \land s \neq s' \supset \text{obscures}(s, s', t)
\]

Proof: Suppose
\[
T_{card_world} \cup NLEA \models (\exists r, s, s', t) \Delta(r, s) \land inclusion(r, t) \land \text{obscures}(s', s, t)
\]

By axiom 7.10 of \( T_{obscures} \) there exists a minimal surface in the \textit{obscures} ordering such that
\[
T_{card_world} \cup NLEA \models (\forall r, s, s', t) \Delta(r, s) \land inclusion(r, t) \land \text{obscures}(s', s, t)
\]
\[
\supset (\exists s'') \text{obscures}(s'', s, t) \land \neg(\exists s^*) \text{obscures}(s^*, s'', t)
\]

Then axiom 7.17 of \( T_{occ} \) entails
\[
T_{card_world} \cup NLEA \models (\forall r, s, s', t) \Delta(r, s) \land inclusion(r, t) \land \text{obscures}(s', s, t)
\]
\[
\land ((\exists s'') \text{obscures}(s'', s, t) \land \neg(\exists s^*) \text{obscures}(s^*, s'', t))
\]
\[
\supset \Delta(r, s'')
\]

which contradicts \( URDA \) and hence contradicts \( NLEA \). \( \Box \)
Chapter 9

Computational Framework for CardWorld

In Part I of this thesis, we presented eight theories of shape, depiction, and occlusion for 2D surfaces. We now want to use this theory for the task of object recognition. In this chapter we provide a framework with which to compute finite models of these axioms. This framework will provide the basis for algorithms and complexity proofs in Part III of the thesis.

Our first step will be to introduce various assumptions that will allow us to simplify the CardWorld axioms for particular reasoning problems. In the preceding chapter we analyzed the occlusion assumption \textit{NLEA}; in this chapter we use it to simplify the occlusion axioms. We also introduce several assumptions pertaining to the shapes of surfaces in the scene; in particular, we will restrict ourselves to convex surfaces and rectangular surfaces.

Given the simplified set of first-order axioms, we then transform them into sets of logically equivalent propositional theories. In the previous chapter we saw how various assumptions entailed by \textit{NLEA} were necessary for this transformation. The resulting set of propositional clauses that use depiction and occlusion literals will be the theory for which the algorithms in following chapters construct models.

We finish the chapter with two algorithms that can be used to satisfy the occlusion clauses in our theory once the depiction clauses have been satisfied. The algorithms in the next chapter will focus on the problem of satisfying the depiction axioms.

9.1 Closure Assumptions

One way in which to simplify a theory is to introduce closure assumptions, which impose additional constraints on either the domains of models or the extensions of relations in the models. For example, in Chapter 3 we introduced closure assumptions over the image domain and the extensions of image relations.

Axiom 3.16 in $T_{\text{scene}}$ can be considered to be a closure assumption over scene objects. Consider for the moment that we did not have this axiom. Although points, edges, and surfaces are scene objects, it
would be consistent for there to exist other kinds of scene objects. By the depiction axioms, it would be consistent for any of these unknown scene objects to be depicted by pixels, junctions, lines, or regions. Further, there are no constraints on multiple depiction for these unknown scene objects, so that a line could depict an unbounded number of them. Thus, there would be an unbounded number of models for any image, and any image could be given the trivial interpretation in which every image object depicts an unknown scene object. Thus axiom 3.16 plays a very important role in the computation of models.

Another closure assumption typically arises in scenes with occluded surfaces. Consider Figure 9.1a and suppose that there is a model $\mathcal{M}$ of $T_{\text{cardworld}} \cup I$ such that

$$\langle r_1, s_1 \rangle, \langle r_2, s_2 \rangle \in \Delta$$

According to the scene axioms in $T_{\text{scene}}$, it is consistent for there to exist surfaces that are not rectangles.
Thus it is consistent for the surface that is occluded to be any of the surfaces in Figure 9.1b.

This also has algorithmic implications. Any image has a trivial interpretation in which the regions depict arbitrary surfaces. There is one interpretation in which each region depicts a unique surface and the surfaces in the scene are all abutting; in other words, the scene is a “mosaic” of arbitrary surfaces [Buffart et al 81]. For example, in Figure 9.2(b), we show the possible surfaces that could be in the image (a).

Also, the notion of grouping regions that depict the same surface depends on the shape of the surfaces in the scene. For example, if we are only considering rectangles, then the regions $r_1$ and $r_2$ in Figure 9.1(c) cannot depict the same surface, but if we also allow l-shapes, then such a grouping would be consistent. Similarly, if we only consider l-shapes, then the regions $r_1$ and $r_2$ in Figure 9.1(d) cannot depict the same surface. In Figure 9.2(a), if there are no restrictions on the shapes of surfaces, then it is consistent for any two nonadjacent regions to depict the same surface. For example, the regions $r_1$ and $r_2$ could depict the surface in Figure 9.2(c). Figure 9.2(d) displays the hidden edges in this case.

Finally, occlusion also depends on the shapes of the surfaces in the scene. In Figure 9.2(a), if there are no constraints on shape, then it is consistent for any surface to occlude any other surface, simply by choosing the appropriate shape. There exists a model $M$ of $T_{cardworld} \cup I$ such that

$$\langle r_1, s_1 \rangle, \langle r_2, s_1 \rangle, \langle r_3, s_3 \rangle \in \Delta \quad \langle s_3, s_1 \rangle \in \text{ontop}$$

if $s_1$ is the surface in Figure 9.2(c), but if all surfaces can only be rectangles, then such occlusion is impossible.

### 9.1.1 Rectangular Shape Closure

We thus propose the notion of shape closure, which will restrict the shapes that appear in the scene. What exactly is a shape? Recall from the axioms in $T_{fg}$ that every surface consists of a set of edges with fixed set of alignment and positional predicates and which has a unique figure ground assignment. A shape will therefore denote a class of surfaces that all obey the same assignments of scene relations. Thus shape closure is a set of expectations about the structure of the world, a way of defining what kinds of surfaces can be present in the scene. In the remainder of the thesis we will restrict our attention to a specific class of surfaces:

**Definition 9.1**

$$\forall s \, \text{rectangle}(s) \equiv$$

$$\left( \exists e_1, e_2, e_3, e_4, v_1, v_2, v_3, v_4 \right) \text{meet}(e_1, e_2, v_1) \land \text{meet}(e_2, e_3, v_2) \land \text{meet}(e_3, e_4, v_3) \land \text{meet}(e_4, e_1, v_4)$$

$$\land \text{parallel}(e_1, e_3) \land \text{parallel}(e_2, e_4) \land \theta(\frac{3\pi}{2}, e_1, e_2) \land \theta(\frac{3\pi}{2}, e_2, e_3) \land \theta(\frac{3\pi}{2}, e_3, e_4) \land \theta(\frac{3\pi}{2}, e_4, e_1)$$

$$\land \text{outside}(e_1, e_2) \land \text{outside}(e_2, e_3) \land \text{outside}(e_3, e_4) \land \text{outside}(e_4, e_1) \land \text{inside}(e_1, e_3) \land \text{inside}(e_2, e_4)$$
Figure 9.2: Examples of interpretations without shape closure.
Thus a rectangle is a surface with exactly four distinct edges such that all edges are either parallel or perpendicular, and there are no holes within the surface. Examples of rectangles are in Figure 9.1(e).

The following assumption states that all surfaces are rectangles, and that all scene elements are either rectangles or part of a rectangle:

**Definition 9.2** The Rectangular Shape Closure Assumption (RSCA) is the sentence

$$\land (((\forall s) \text{ surface}(s) \equiv \text{rectangle}(s))$$

$$\land (((\forall z) \text{ scene.object}(z) \supset \text{rectangle}(z) \lor (\exists y) \text{rectangle}(y) \land \text{part}(z,y))$$

$$\land (((\forall z) \text{ rectangle}(z) \land \text{part}(z,z) \supset y = z)))$$

Notice that RSCA is restricted to a unique shape, namely rectangles. In general, shape assumptions can be used to define arbitrary surfaces, and can be relaxed to include multiple shapes. All of these forms of shape closure are closely related to the libraries of objects used in model-based vision, in which the shape closure is assumed a priori.

We can also specify closure assumptions which are entailed by shape closure.

**Definition 9.3** The rectangular alignment relation closure assumption (RACA) is the sentence

$$\forall s, e, e' \land \text{part}(e, s) \land \text{part}(e', s) \supset \text{parallel}(e, e') \lor \theta(\frac{\pi}{2}, e, e') \lor \theta(\frac{3\pi}{2}, e, e')$$

i.e., all edges in a surface are either parallel or they form right angles.

For clarity, we will refer to $\theta(\frac{\pi}{2}, e, e')$ or $\theta(\frac{3\pi}{2}, e, e')$ as $\text{perp}(e, e')$, so that RACA can be expressed as

$$\forall s, e, e' \land \text{part}(e, s) \land \text{part}(e', s) \supset \text{parallel}(e, e') \lor \text{perp}(e, e')$$

It is easy to see that this assumption is entailed by the rectangular shape closure assumption and the definition of rectangle:

**Proposition 9.1**

$$T_{\text{cardworld}} \models \text{RSCA} \supset \text{RACA}$$

We can also define conditions on the extensions of the figure/ground relations for surfaces:
Definition 9.4 The rectangular figure ground closure assumption (RFGC) is the sentence

\[(\forall s, e, e', a) \text{ part}(e, s) \wedge \text{ part}(e', s) \supset \text{ inside}(e, e') \equiv \theta(a, e, e') \wedge a < \pi)\]

i.e., all edges in a surface are convex 1.

Again, it is also easy to see that this assumption is entailed by the rectangular shape closure assumption:

Proposition 9.2

\[T_{\text{cardworld}} \models \text{RSCA} \supset \text{RFGC}\]

Thus closure on rectangular shapes entails closure of the various scene relations that define the shape: these subordinate assumptions will be more useful than the original RSCA. It should be emphasized that RACA and RFGC are defined relative to the RSCA: if RSCA entailed the existence of other different shapes, the RACA and RFGC would also change correspondingly.

In addition to the scene relations of \(\theta\) and inside, the notion of shape also has a geometric aspect, which specifies the position of points in the surface. Consider the rectangle axiom: the lengths of the edges are not specified in the axiom and can be considered as parameters. In this work we will restrict these lengths to be fixed constants; this will make it easier to compute the position of points in a surface, which we will need for the occlusion axioms (recall that interior and boundary use geometric relations that require us to know the position of points).

Definition 9.5 The parametrized shape closure assumption (PSCA) used in this thesis is the following sentence:

\[(\forall s) \text{ rectangle}(s) \supset (\exists e_1, e_2, e_3, e_4) \text{ edge}(e_1) \wedge \text{ edge}(e_2) \wedge \text{ edge}(e_3) \wedge \text{ edge}(e_4)\]

\[\wedge \text{ part}(e_1, s) \wedge \text{ part}(e_2, s) \wedge \text{ part}(e_3, s) \wedge \text{ part}(e_4, s) \wedge \text{ parallel}(e_1, e_3) \wedge \text{ parallel}(e_2, e_4)\]

\[\wedge \text{ length}(e_1, D_1) \wedge \text{ length}(e_2, D_2) \wedge \text{ length}(e_3, D_1) \wedge \text{ length}(e_4, D_2)\]

where \(D_1\) and \(D_2\) are fixed constants.

Note that all the rectangles in Figure 9.1(e) have edges with different lengths.

Given the positions of some subset of points in a surface, we can use PSCA to determine the sets of possible positions for all other points in the surface using Hilbert's model of geometry, in which edges of a surface are represented as segments in a polygon. This geometric component of PSCA is equivalent to a set of algebraic constraints on the positions of all points that are part of a surface. In general, these constraints depend on the meet and polygonal relations, which state how the edges of a surface are connected, as well as the lengths of the edges in a surface. This thesis will not be concerned with these geometric constraints; we will only require that they satisfy one property - once the positions of the vertices have been fixed, the constraints yield the positions of all other points in the surface.

---

1 Recall the definition of convex from Chapter 4.
9.1.2 Assumptions for Convex Surfaces

We can generalize many results about rectangles to the class of surfaces which intuitively are convex. In this section, we consider two assumptions about convex surfaces using $T_{\text{cardworld}}$, and investigate how theories which are restricted to this class of surfaces can be used to strengthen results about depiction and occlusion.

We first consider a characterization of convex surfaces with respect to figure/ground assignments:

**Definition 9.6** The Convex Figure/Ground Assumption (CFG) is the sentence

\[(\forall s, e_1, e_2, a) \text{part}(e_1, s) \land \text{part}(e_2, s) \supset (\text{inside}(e_1, e_2) \equiv \theta(a, e_1, e_2) \land a < \pi) \land (\text{outside}(e_1, e_2) \equiv \theta(a, e_1, e_2) \land a > \pi)\]

For any convex surface, CFGA provides a complete figure/ground assignment based on the alignment relations between the edges in the surface, so that the axioms in $T_{fg}$ are satisfied. The surface in Figure 4.2(a) is convex, but the surfaces in Figure 4.2(b),(c) are not convex; in each case there exists a model $\mathcal{M}$ of $T_{\text{cardworld}} \cup \text{CFG}A$ such that

\[(e_3, e_4) \in \text{inside}, (a, e_3, e_4) \in \theta, a > \pi\]

It is easy to see that this assumption is already satisfied if we are making the rectangular shape closure assumption:

**Proposition 9.3**

\[T_{\text{cardworld}} \models \text{RFGA} \implies \text{CFG}A\]

**Proposition 9.4**

\[T_{\text{cardworld}} \cup \text{RSCA} \models \text{CFG}A \equiv \text{RFGC}\]

We can also introduce an additional assumption which characterizes convex surfaces with respect to interior:

**Definition 9.7** The Convex Surfaces Assumption (CSA) is the sentence

\[(\forall e, s, t) \text{boundary}(e, t) \equiv (\text{part}(e, s) \land \text{interior}(s, t))\]

In other words, all edges in a convex surface are boundary edges for any position interior to the surface. This is similar to the traditional definition of convex surfaces ([de Berg et al 97]).

### 9.2 Simplifying the CardWorld Axioms

In this section, we will use RSCA (rectangular shape closure assumptions), the convex surface assumptions, and NLEA to simplify the axioms of CardWorld. The objective is to show that either axioms in $T_{\text{cardworld}}$ are entailed by the assumptions that we are making, or to show that axioms in $T_{\text{cardworld}}$ can be strengthened by our assumptions.

Let $T_{cw}$ the axioms in Figure 9.3. The central result of this section is that $T_{cw}$ is logically equivalent to $T_{\text{cardworld}}$ in conjunction with the shape closure assumptions and the occlusion assumption NLEA:
CHAPTER 9. COMPUTATIONAL FRAMEWORK FOR CardWorld

\[ RSCA \] (9.1)
\[ PSCA \] (9.2)
\[ RACA \] (9.3)

\[
(\forall e, e', l, l', s, a) \Delta(l, e) \land \Delta(l', e') \land \text{part}(e, s) \land \text{part}(e', s) \supset (\theta_i(l, l') \equiv \theta(a, e, e'))
\] (9.4)

\[
(\forall e, e', l, l', s) \Delta(l, e) \land \Delta(l', e') \land \text{part}(e, s) \land \text{part}(e', s) \supset (\text{parallel}(l, l') \equiv ||(e, e'))
\] (9.5)

\[
(\forall e, e', l, l', r, a) \text{part}(e, s) \land \text{part}(e', s) \land \Delta(l, e) \land \Delta(l', e') \land \text{in}(l, r) \land \theta(a, l, l') \supset ((\Delta(r, s) \equiv \text{between}(l, l', r)) \equiv a < \pi)
\] (9.6)

\[
(\forall l, l', r, r', e, s) \land \text{in}(l, r) \land \text{in}(l', r') \land \Delta(l, e) \land \Delta(l', e) \land \text{part}(e, s) \supset
\]
\[
[(\Delta(r, s) \equiv \Delta(r', s)) \equiv (\text{between}(l, l', r) \equiv \text{between}(l', l', r'))]
\] (9.7)

\[
(\forall l_1, l_2, e_1, e_2, s) \Delta(l_1, e_1) \land \Delta(l_2, e_2) \land \text{part}(e_1, s_1) \land \text{part}(e_2, s_1) \supset (e_1 = e_2 \equiv \theta_i(\pi, l_1, l_2))
\] (9.8)

\[
(\exists e, p, t) \text{boundary}(e, t) \land \text{part}(e, s) \land \text{part}(p, e) \land \text{position}(p, t) \land \text{position}(s, t)
\] (9.9)

\[
(\forall s, s', t) \text{interior}(s, t) \land \text{interior}(s', t) \equiv
\]
\[
(\text{obsures}(s, s', t) \lor \text{obsures}(s', s, t) \lor (\exists e') \text{part}(e, s) \land \text{part}(e', s') \land \text{abuts}(e, e', t))
\] (9.10)

\[
(\forall e_1, e_2, s_1, s_2, s_3, t) \text{obsures}(s_1, s_2, t) \land \text{abuts}(e_1, e_2, t) \land \text{part}(e_1, s_2) \land \text{part}(e_2, s_3) \supset \text{obsures}(s_1, s_3, t)
\] (9.11)

\[
(\forall e_1, e_2, t) \text{abuts}(e_1, e_2, t) \supset \text{abuts}(e_2, e_1, t)
\] (9.12)

\[
(\forall s_1, s_2, t_1, t_2) \text{obsures}(s_1, s_2, t_1) \land \text{interior}(s_1, t_1) \land \text{interior}(s_2, t_2) \supset \text{obsures}(s_1, s_2, t_2)
\] (9.13)

\[
(\forall r, s) ((\exists t) \text{interior}(s, t) \land \text{inclusion}(r, t)) \equiv (\Delta(r, s) \lor (\exists s') \Delta(r, s') \land \text{ontop}(s', s))
\] (9.14)

\[
(\forall l, e, s, r, r') \Delta(l, e) \land \text{part}(e, s) \land \text{in}(l, r) \land \text{in}(l, r') \land r \neq r' \supset (\Delta(r, s) \lor \Delta(r', s))
\] (9.15)

\[ CFGA \] (9.16)
\[ CSA \] (9.17)

\[
(\forall p, q, t) \text{point}(p) \land \Delta(q, p) \supset (\text{position}(p, t) \equiv \text{position}(q, t))
\] (9.18)

\[
(\forall s, s') \text{ontop}(s, s') \supset \text{ontop}(s', s)
\] (9.19)

\[
(\forall s, s') \text{abutting}(s, s') \supset \text{ontop}(s', s)
\] (9.20)

\[
(\forall s) \text{notabutting}(s, s)
\] (9.21)

\[ NSEA \] (9.22)
\[ IDA \] (9.23)
\[ OSDA \] (9.24)
\[ URDA \] (9.25)
\[ DTR \] (9.26)

Figure 9.3: The simplified CardWorld axioms $T_{cw}$ (excluding conservative definitions).
Theorem 9.1 For any planar image $I$, $T_{cw} \cup I$ is logically equivalent to $T_{\text{cardworld}} \cup I \cup \text{NLEA} \cup \text{RACA} \cup \text{RFGA} \cup \text{CSA}$.

Proof: All axioms in $T_{cw}$ (except for axiom 9.14) are also axioms in $T_{\text{cardworld}}$. We therefore only need to show that axiom 9.14 is entailed by $T_{\text{cardworld}} \cup I \cup \text{NLEA}$, and we will have shown

$$T_{\text{cardworld}} \cup I \cup \text{NLEA} \cup \text{RACA} \cup \text{RFGA} \cup \text{CSA} \models T_{cw}.$$

Lemma 9.1 For any planar image $I$,

$$T_{\text{cardworld}} \cup I \cup \text{NLEA} \models (\forall r,s) ((\exists t) \text{interior}(s,t) \land \text{inclusion}(r,t)) \equiv \Delta(r,s) \lor (\exists s') \Delta(r,s') \land \text{ontop}(s',s)$$

Proof: Axiom 6.11 of $T_{\text{interior}}$ gives

$$T_{\text{cardworld}} \cup I \cup \text{NLEA} \models (\forall r,s_1,s_2) ((\Delta(r,s_1) \lor (\Delta(r,s_2) \land \text{ontop}(s_2,s_1))))$$

$$\supset (\exists t) \text{interior}(s_1,t) \land \text{inclusion}(r,t)$$

For the converse, axioms 7.17, 6.12, and URDA give us:

$$T_{\text{cardworld}} \cup I \cup \text{NLEA} \models (\forall r,s_1,t) \text{interior}(s_1,t) \land \text{inclusion}(r,t) \supset$$

$$(\Delta(r,s_1) \land \neg(\exists s') \text{obsures}(s',s_1,t))$$

$$\lor (\exists s_2) (\text{interior}(s_2,t) \land \text{obsures}(s_2,s_1,t) \land \Delta(r,s_2))$$

which entails

$$T_{\text{cardworld}} \cup I \cup \text{NLEA} \models (\forall r,s_1,t) \text{interior}(s_1,t) \land \text{inclusion}(r,t) \supset$$

$$(\Delta(r,s_1) \land \neg(\exists s') \text{obsures}(s',s_1,t))$$

$$\lor (\exists s_2) \text{ontop}(s_2,s_1) \land \Delta(r,s_2)$$

$\square$

For the other direction of the main theorem, the following lemmas show how each of the sentences in each module in $T_{\text{cardworld}}$ are entailed by $T_{cw}$. $\square$

Simplifying $T_{\text{scene}}$

Lemma 9.2

$$T_{cw} \models T_{\text{scene}}$$

Proof: RSCA is a sentence in $T_{cw}$, and it is easy to see that

$$\text{RSCA} \models T_{\text{scene}}$$

since all scene objects are either a unique rectangular surface or they are part of a unique rectangular surface. $\square$
Simplifying $T_{fg}$

**Lemma 9.3**

$T_{cw} \models T_{fg}$

**Proof:** $RSC$ is a sentence in $T_{cw}$, and it is easy to see that $RSC$ entails axioms 4.1-4.7 of $T_{fg}$ from the definition of rectangle, which completely specifies the extensions of meet, connected, outer, and the polygonal relations for the four edges in a rectangular surface.

We therefore need to show how $T_{cw}$ entails the figure/ground axioms in $T_{fg}$.

**Lemma 9.4**

$T_{fg} \cup CFGA \models (\forall s, e, e') part(e, s) \land part(e', s) \supset \neg indet(e, e')$

**Proof:** Suppose $\mathcal{M}$ is a model of $T_{fg} \cup CFGA$ such that for any variable assignment

$\mathcal{M}, \sigma \models indet(e, e') \land \theta(a, e, e') \land \theta(\pi - a, e', e')$

If $a < \pi$, then we have

$\mathcal{M}, \sigma \models inside(e, e') \land \neg outside(e, e')$

and if $a > \pi$, then we have

$\mathcal{M}, \sigma \models \neg inside(e, e') \land outside(e, e')$

both of which contradict $T_{fg}$. □

**Lemma 9.5**

$T_{fg} \cup CFGA \models (\forall s, e) part(e, s) \supset outer(e)$

**Proof:** Suppose $\mathcal{M}$ is a model of $T_{fg} \cup CFGA$ such that for any variable assignment

$\mathcal{M}, \sigma \models \neg outer(e)$

By $T_{fg}$, there exists an edge $e_1$ such that

$\mathcal{M}, \sigma \models \neg outer(e_1) \land connected(e, e_1) \land (inside(e, e_1) \equiv (\exists a) \theta(a, e, e_1) \land a > \pi)$

which violates $CFGA$. □

Using these lemmas, we can see that the figure/ground axioms of $T_{fg}$ are satisfied by $CFGA$. □

Simplifying $T_\Delta$

**Lemma 9.6**

$T_{cw} \models T_\Delta$

**Proof:** Since there are no indet edges in a convex surface, axioms 5.5 and 5.6 of $T_\Delta$ are trivially satisfied.

The remaining depiction axioms in $T_\Delta$ simplify to axioms 9.4-9.7 in $T_{cw}$ (see Figure 9.3).

Finally, axiom 5.4 of $T_\Delta$ can be strengthened to the following:

**Lemma 9.7**

$T_{cardworld} \cup CFGA \models (\forall l_1, l_2, e_1, e_2, s) \Delta(l_1, e_1) \land \Delta(l_2, e_2) \land part(e_1, s_1) \land part(e_1, s_1) \supset \Delta(e_1 = e_2 = 1 \equiv \theta^i(\pi, l_1, l_2))$

**Proof:** By the definition of $CFGA$, there cannot be edges in a convex surface which have relative alignment of $\pi$. Since the collinear lines must depict edges in the same surface, the only possibility is that the collinear lines must depict the same edge. □

□
Simplifying $T_{interior}$ Since the figure-ground relations are fixed by $CFG_A$, we can simplify the definition of boundary edges so that it refers only to geometric relations:

**Proposition 9.5**

$$T_{cw} \cup CFG_A \models (\forall e, s, t) \text{part}(e, s) \supset \text{boundary}(e, t) \equiv (\exists e_1) \text{part}(e_1, s) \wedge \text{sweep}(e, e_1, t)$$

$$\wedge ((\forall p, t') \text{part}(p, e) \wedge \text{position}(p, t') \supset \text{noncrossing}(s, t, t'))$$

Further, since there are no outer edges in a convex surface, the only axiom in $T_{interior}$ which is not trivially satisfied is axiom 6.7 of $T_{interior}$, which appears in $T_{cw}$.

Simplifying $T_{obscures}$ Using $CS_A$, we can simplify the occlusion axioms in $T_{obscures}$ for convex surfaces, particularly axiom 7.8, which forbids interpenetrating surfaces.

**Lemma 9.8**

$$T_{cw} \models T_{obscures}$$

**Proof:** We proceed by showing how $T_{cw}$ entails axioms within $T_{obscures}$.

We begin by observing that the axiom which states that the extension of $obscures$ is the same for all positions in the same convex components of the surfaces is simplified if we restrict ourselves to convex surfaces:

**Lemma 9.9**

$$T_{cw} \cup CS_A \models ((\forall s_1, s_2, t_1, t_2) \text{obscures}(s_1, s_2, t_1) \wedge \text{interior}(s_1, t_2) \wedge \text{interior}(s_2, t_2)$$

$$\wedge \text{noncrossing}(s_1, t_1, t_2) \wedge \text{noncrossing}(s_2, t_1, t_2) \supset \text{obscures}(s_1, s_2, t_2))$$

$$\equiv ((\forall s_1, s_2, t_1, t_2) \text{obscures}(s_1, s_2, t_1) \wedge \text{interior}(s_1, t_2) \wedge \text{interior}(s_2, t_2) \supset \text{obscures}(s_1, s_2, t_2))$$

Using this property, we can simplify the axioms which specify the relationship between $obscures$ and $abuts$. First, we will introduce a new binary relation over surfaces based on the $abuts$ relation:

**Definition 9.8**

$$(\forall s, s') \text{abutting}(s, s') \equiv (\exists e, e', t) \text{part}(e, s) \wedge \text{part}(e', s') \wedge \text{abuts}(e, e', t)$$

Thus, two surfaces are abutting iff they contain edges which satisfy the $abuts$ relation.

We can show that axioms using the $obscures$ and $abuts$ relations (which contain positions as arguments) are equivalent to axioms using $ontop$ and $abutting$ (which contain only surfaces as arguments).

**Lemma 9.10**

$$T_{cw} \cup CS_A \models (\forall s_1, s_2, t) \text{obscures}(s_1, s_2, t) \supset \neg \text{obscures}(s_2, s_1, t))$$

$$\equiv ((\forall s, s') \text{ontop}(s, s') \supset \neg \text{ontop}(s', s))$$

$$T_{cw} \cup CS_A \models (\forall e_1, e_2, s_1, s_2, t) \text{abuts}(e_1, e_2, t) \wedge \text{part}(e_1, s_1) \wedge \text{part}(e_2, s_2) \supset \neg \text{obscures}(s_1, s_2, t))$$

$$\equiv ((\forall s, s') \text{ontop}(s, s') \supset \neg \text{abutting}(s, s'))$$
Proof: We can rewrite axiom 7.8 of $T_{obsn<res}$ as

$$(\forall s, s', t_1, t_2) \neg \text{obscures}(s', s, t_1) \land \text{noncrossing}(s, t_1, t_2) \land \text{noncrossing}(s', t_1, t_2)$$

$$\land \text{interior}(s, t_2) \land \text{interior}(s', t_2) \supset \neg \text{obscures}(s', s, s, t_2)$$

which by axiom 7.1 of $T_{obsn<res}$ is equivalent to

$$(\forall s, s', t_1, t_2) ((\text{obscures}(s, s', t_1) \lor (\exists e, e') \text{abuts}(e, e', t_1)) \land \text{part}(e, s') \land \text{part}(e', s'))$$

$$\land \text{noncrossing}(s, t_1, t_2) \land \text{noncrossing}(s', t_1, t_2) \land \text{interior}(s, t_2) \land \text{interior}(s', t_2) \supset \neg \text{obscures}(s', s, s, t_2)$$

which by the definitions of $\text{ontop}$ and $\text{abutting}$ and is equivalent to the two sentences

$$(\forall s, s') \text{ontop}(s, s') \supset \neg \text{ontop}(s', s)$$

and

$$(\forall s, s') \text{abutting}(s, s') \supset \neg \text{ontop}(s', s)$$

□

With nonconvex surfaces, we can have scenes with mutually occluding surfaces (i.e. surfaces that satisfy $\text{ontop}(s, s') \land \text{ontop}(s', s)$), so that this equivalence of $T_{cw}$ and $T_{obsn<res}$ can fail if we allow nonconvex surfaces.

By the definition of $\text{abutting}$, it is easy to see that axiom 7.7 of $T_{obsn<res}$

$$(\forall e_1, e_2, s, t) \text{abuts}(e_1, e_2, t) \land \text{part}(e_1, s) \supset \neg \text{part}(e_2, s)$$

is equivalent to the sentence

$$(\forall s) \neg \text{abutting}(s, s)$$

The uniquely depicting region assumption can guarantee that there always exists a minimal surface in the $\text{obsn<res}$ ordering at any position, so that axiom 7.10 of $T_{obsn<res}$ is trivially satisfied:

Lemma 9.11

$T_{kernel} \cup DKRE \cup URDA \models (\forall s_1, t) \text{interior}(s_1, t) \supset (\exists s_2)(\forall s_2) \text{interior}(s_2, t) \land \neg \text{obscures}(s_3, s_2, t)$

Proof: By closure over image objects, the number of image objects is fixed for a given image $I$. Since $NLEA$ entails $URDA$, a region can depict at most one surface; by scene object closure only surfaces are depicted by regions; by $NLEA$ and all unoccluded surfaces must be depicted by regions. Putting all of this together we see that if there are $|R|$ regions in the image, there are at most $|R|$ surfaces in the scene. By the shape closure assumption, all surfaces are rectangles, which have a fixed number of component scene objects. Further, all edges and vertices must be part of a surface, so that the number of scene objects is bounded by a linear function of the number of surfaces.

Since there is a bounded finite number of surfaces in a model of $T_{kernel} \cup DKRE \cup URDA$, there cannot be an unbounded $\text{obsn<res}$ ordering, and hence axiom 7.10 of $T_{obsn<res}$ is trivially satisfied. □

Note that any set of assumptions that imposes a bound on the number of depicted surfaces in a model will also entail axiom 7.10 of $T_{obsn<res}$. A similar argument can be used to show that in any finite domain, the $\text{obsn<res}$ ordering must be discrete, so that axiom 7.9 is also trivially satisfied.

□
Simplifying $T^\Delta_{\text{interior}}$ The occlusion assumption $NLE.A$ allows us to simplify the axioms of $T^\Delta_{\text{interior}}$: 

**Lemma 9.12** For any planar image $I$,

$$T_{cw} \cup I \models T^\Delta_{\text{interior}}$$

**Proof:** Several axioms of $T^\Delta_{\text{interior}}$ are entailed by one of the depiction assumptions related to $NLE.A$:

**Lemma 9.13** For any planar image $I$,

$$T_{cw} \cup I \models (\forall r, s_1, t) \text{interior}(s_1, t) \land \text{inclusion}(r, t) \supset (\exists s_2) \Delta(r, s_2) \land \text{interior}(s_2, t)$$

**Proof:** This follows from Lemma 9.1. □

**Lemma 9.14**

$$T_{cw} \models (\forall r, s, t) \Delta(r, s) \supset [(\text{inclusion}(r, t) \equiv \text{interior}(s, t)) \supset \exists l \text{border}(l, r) \supset (\exists s) \text{boundary}(l, s) \land \text{part}(e, s))$$

**Proof:** This follows from $IND.A$. □

Since all edges in a convex surface must be outer edges, there can be no holes in a convex surface: thus the axiom 6.14 of $T^\Delta_{\text{interior}}$ is trivially satisfied by $NLE.A$ and $CFG.A \cup CS.A$.

Thus, all of the axioms of $T^\Delta_{\text{interior}}$ are either explicitly axioms of $T_{cw}$ or they are entailed by $T_{cw}$. □

**Simplifying $T_{\text{occ}}$** We first characterize the relationship between $T_{cw}$ and $NLE.A$:

**Lemma 9.15** For any planar image $I$,

$$T_{cw} \cup I \models NLE.A$$

**Proof:** Notice that the occlusion assumption $NLE.A$ is not explicitly included in $T_{cw}$: rather, we use the depiction assumptions which are equivalent to $NLE.A$. In Chapter 8, we showed that $NLE.A$ is equivalent to the conjunction of $BREA$ and $IND.A$.

We showed that $BREA$ is equivalent to $IDA$ and $OSDA$, which are included as axioms in $T_{cw}$.

We also showed that $IND.A$ is equivalent to the conjunction of $URDA$, $WDTR$, and $HREA$. The first two assumptions are included as axioms within $T_{cw}$. Since all edges in a convex surface must be outer edges, there can be no holes in a convex surface; thus the depiction assumption $HREA$ are trivially satisfied by $NLE.A$ and $CFG.A \cup CS.A$. □

The occlusion assumption $NLE.A$ (and the depiction assumptions that it entails) will allow us to also simplify the axioms of $T_{\text{occ}}$:

**Lemma 9.16** For any planar image $I$,

$$T_{cw} \cup I \models T_{\text{occ}}$$
CHAPTER 9. COMPUTATIONAL FRAMEWORK FOR CARDWORLD

Proof: It is trivial to see that for any planar image $I$, $NLEA$ entails the following axiom of $T_{occ}$:

$$T_{cw} \cup I \models (\forall p, s, t, q) \text{part}(p, s) \land \text{position}(p, t) \supset \Delta(q, p) \equiv (\exists s') \text{obscures}(s', s, t)$$

and

$$T_{cw} \cup I \models (\forall l, e, s) \Delta(l, e) \land \text{part}(e, s) \supset (\exists p, t) \text{part}(p, e) \land \text{position}(p, t) \land (\exists s') \text{obscures}(s', s, t)$$

We also have

$$T_{cw} \cup I \models (\forall s, r) \text{surface}(s) \supset \Delta(r, s) \equiv (\exists t) \text{interior}(s, t) \land \text{inclusion}(r, t) \land (\exists s') \text{obscures}(s', s, t)$$

by $URDA$, $INDA$, and axiom 9.14 of $T_{cw}$.

$\square$

9.3 A Clausal Formulation

In this section, we present a framework for computing models of $T_{cw} \cup I$ for some planar image $I$ by specifying sets of clauses that are logically equivalent to $T_{cw} \cup I$. We will use this theory in algorithms to construct models satisfying the depiction axioms and also to prove properties about constructs used in NP-completeness theorems of the next chapter.

We will first show that the occlusion assumption $NLEA$, together with the shape closure assumptions, allows us to place a finite bound on the domain of a model of $T_{cardworld}$. We then specify the propositional theory which is equivalent to $T_{cw}$ under these finite model assumptions. This propositional theory contains two types of sentences – depiction sentences and occlusion sentences. We therefore introduce the notion of poses for surfaces, which form the basis for specifying the occlusion sentences in the propositional theory.

9.3.1 Finite Model Assumptions

Using the results of the previous two chapters, recall that there is an upper bound on the number of surfaces depicted in an image, namely the number of regions in the image. Thus if we fix the number of depicted surfaces, the depiction axioms become essentially propositional.

We begin by demonstrating that with the assumptions that we have considered to this point, we can guarantee that all models of our theory will be finite.

**Theorem 9.2** All models in the depicted kernel of $T_{cw} \cup I$ have finite bounded domains.

**Proof:** By closure over image objects, the number of image objects is fixed for a given image $I$. Since $NLEA$ entails $URDA$, a region can depict at most one surface; by scene object closure only surfaces are depicted by regions; by $NLEA$ and $DKA$ all surfaces must be depicted by regions. Putting all of this together we see that if there are $|R|$ regions in the image, there are at most $|R|$ surfaces in the scene. By the shape closure assumption, all surfaces are rectangles, which
have a fixed number of component scene objects \(^2\). Further, all edges and vertices must be part of a surface, so that the number of scene objects is bounded by a linear function of the number of surfaces. \(\Box\)

We thus have a finite characterization of the set of models for an image. All that is required to construct a model of \(\text{CardWorld} \cup I\) is to determine the depiction literals satisfied by the model: the number of elements in the domain is finite. This has important algorithmic implications. It allows us to construct kernel models in a top-down fashion, since we can specify the domain by instantiating the \(n\) surfaces in the kernel, along with the scene objects that are parts of these surfaces. (Recall that the number of surfaces is bounded by the number of regions in the image.) We then satisfy the depiction axioms by matching the scene objects to image objects.

More formally, for a given image, we can define the following set of scene objects:

**Definition 9.9** Given an image \(I\), \(\text{Rec}(I)\) is the set of scene elements consisting of rectangles and parts of rectangles, such that there is a 1-1 mapping between the surfaces in the set and the regions in the image \(I\).

The advantage of this top-down approach is that it allows us to use a Herbrand domain for the models and also to make the unique names assumption for scene objects. This follows since every scene object instantiated from the surface axioms is distinct, as well as the fact that every point and edge is in a unique surface. This requires that we extend the language so that we have constants denoting all of the scene objects in \(\text{Rec}(I)\).

**Definition 9.10** Given an image \(I\), \(\mathcal{L}(\text{Rec}(I))\) is the expansion of \(\mathcal{L}_{\text{cardworld}}\) by adding a set of new constants such that each constant denotes a unique scene object in \(\text{Rec}(I)\).

Essentially, we are using the definition of rectangle to specify the domain: \(\mathcal{L}(\text{Rec}(I))\) extends the language of the original theory by introducing constants for each object in the domain.

For example, consider the image \(I\) in Figure 9.1a. The new constants in \(\mathcal{L}(\text{Rec}(I))\) denoting the scene objects are

\[s_1, s_2, e_1, e_2, e_3, e_4, e_5, e_6, e_7, e_8, v_1, v_2, v_3, v_4, v_5, v_6, v_7, v_8\]

To construct a model of \(\text{CardWorld} \cup I\), we need only match the depicted scene objects with the image objects in \(I\).

For all images in the examples in the remainder of this thesis, we will use the language \(\mathcal{L}(\text{Rec}(I))\).

### 9.3.2 A Useful Depiction Assumption

Throughout the remaining work, we will be using relations among lines (such as parallel and perp) to conjecture certain scene and depiction relations, particularly for relations between the depiction of lines and the depiction of regions. Thus we will be using the following:

\(^2\)This normally holds for arbitrary surfaces, since in practice any surface has a finite number of component scene objects. However, \(\text{RSCA}\) is the only axiom that allows us to say this, since no other axiom precludes surfaces with an infinite number of edges. Of course, closure assumptions for other classes of surfaces may be introduced if necessary.
Figure 9.4: Example image for the Region Depiction Assumption (RDA).

Definition 9.11 The Region Depiction Assumption (RDA) is the sentence

$$(\forall r, s) \, \text{surface}(s) \land \Delta(r, s) \supset (\exists l, e) \, \text{in}(l, r) \land \Delta(l, e) \land \text{part}(e, s)$$

This assumption states that every region that depicts a surface contains a line that depicts an edge in that surface, and is made in [Jacobs 88].

In models of $T_{\text{cardworld}} \cup DK A$ which violate RDA, there exist surfaces which are depicted, yet which have no depicted edges.

Example: Let $I$ be the image in Figure 9.4.

There exists a model $\mathcal{M}$ of $T_{\text{cardworld}} \cup I$ such that

$$\langle r_1, s_1 \rangle, \langle r_2, s_2 \rangle, \langle r_3, s_3 \rangle, \langle r_4, s_4 \rangle, \langle l_1, e_1 \rangle, \langle l_2, e_2 \rangle, \langle l_3, e_3 \rangle \in \Delta$$

$$\langle e_1, s_1 \rangle, \langle e_2, s_2 \rangle, \langle e_3, s_3 \rangle \in \text{part}$$

This model does not satisfy RDA, since the surface $s_4$ is depicted, yet none of its edges are depicted.

There exists another model $\mathcal{N}$ of $T_{\text{cardworld}} \cup I$ such that

$$\langle r_1, s_1 \rangle, \langle r_2, s_2 \rangle, \langle r_3, s_3 \rangle, \langle r_4, s_4 \rangle, \langle l_1, e_1 \rangle, \langle l_2, e_2 \rangle, \langle l_3, e_3 \rangle, \langle l_3, e_4 \rangle \in \Delta$$

$$\langle e_1, s_1 \rangle, \langle e_2, s_2 \rangle, \langle e_3, s_3 \rangle, \langle e_4, s_4 \rangle \in \text{part}$$

This model satisfies RDA, since each depicted surface has a depicted edge.
Finally, there is another model $B$ of $T_{cardworld} \cup I$ such that

$$\langle r_1, s_1 \rangle, \langle r_2, s_2 \rangle, \langle r_3, s_3 \rangle, \langle l_1, e_1 \rangle, \langle l_2, e_2 \rangle, \langle l_3, e_3 \rangle \in \Delta$$

$$\langle e_1, s_1 \rangle, \langle e_2, s_2 \rangle, \langle e_3, s_3 \rangle \in \text{part}$$

This model also satisfies $RDA$, since each depicted surface has a depicted edge. In this model, the region $r_4$ does not depict any surface. □

### 9.3.3 Poses for Occluded Surfaces

Before specifying the propositional theory which is equivalent to $T_{cw} \cup I$, we need to revisit the relationship between occlusion and depiction. In order to specify the class of sentences in the propositional theory which correspond to occlusion sentences, it will be useful to introduce the concept of the *pose* of a surface (c.f. [Grimson 90]):

**Definition 9.12** Let $M$ be a model of $T_{cardworld} \cup I \cup NLEA$. A pose of a depicted surface $s$ is a conjunction of *ground ontop* literals and *ground surface depiction* literals such that

$$M \models \text{ontop}(s, s) \land \Delta(r, s)$$

Thus, a pose for a surface $s$ can be used to identify the set of depicted surfaces which occlude $s$.

**Example:** Let $I$ be the image in Figure 9.5a, and let $L = L(I) \cup \{s_1, s_2, s_3, s_4, s_5\}$. There is a model $M_C$ of $T_{cardworld} \cup NLEA \cup I$

$$M_C \models \Delta(r_1, s_1) \land \Delta(r_2, s_2) \land \Delta(r_3, s_3) \land \Delta(r_4, s_4) \land \Delta(r_5, s_5)$$

In this model the surface denoted by $s_1$ has the unique pose

$$\text{ontop}(s_2, s_1) \land \text{ontop}(s_5, s_1)$$

The surface denoted by $s_2$ has the unique pose

$$\text{ontop}(s_5, s_2)$$

The surface denoted by $s_3$ has the unique pose

$$\text{ontop}(s_1, s_3)$$

and the surface denoted by $s_4$ has the unique pose

$$\text{ontop}(s_1, s_4)$$

The surface denoted by $s_5$ has an empty pose since all parts of this surface is depicted. □
Figure 9.5: Examples of poses
Surfaces need not have the same pose in all models of an image. Note that the pose of a surface exists for a given set of undepicted scene objects and \textit{interior} depends on the positions of all points in the surface. In any given model for an image, the positions of all points are fixed. However, if these positions are not fixed by the theory, then we will have multiple models, and in each model the surfaces will have different poses.

Example: Consider the image \( I \) in Figure 9.5(b). Let \( \mathcal{L} = L(I) \cup \{s_1, s_2, s_3, s_4, s_5\} \). There are two models \( \mathcal{M}_1 \) and \( \mathcal{M}_2 \) of \( T_{\text{cardworld}} \cup \text{NLEA} \cup I \) such that
\[
\mathcal{M}_1 = \Delta(r_1, s_1) \land \Delta(r_2, s_2) \land \Delta(r_3, s_3) \land \Delta(r_4, s_4) \land \Delta(r_5, s_5)
\land \text{ontop}(s_2, s_1) \land \text{ontop}(s_4, s_1) \land \text{ontop}(s_5, s_1)
\]
and
\[
\mathcal{M}_2 = \Delta(r_1, s_1) \land \Delta(r_2, s_2) \land \Delta(r_3, s_3) \land \Delta(r_4, s_4) \land \Delta(r_5, s_5)
\land \text{ontop}(s_3, s_1) \land \text{ontop}(s_4, s_1) \land \text{ontop}(s_5, s_1)
\]
These poses are represented showing the hidden occluded edges in \( s_1 \) in Figure 9.5(c) and (d). \( \Box \)

The notion of pose tells us \textit{which} surfaces occlude an undepicted scene object. For example, in Figure 9.5(e) there is a model \( \mathcal{M} \) such that
\[
\mathcal{M} = \text{part}(e_1, s_1) \land \text{position}(p_1, t_1) \land \neg(\exists q) \Delta(q, p_1) \land \text{part}(p_2, s_1) \land \text{position}(p_2, t_2) \land \neg(\exists q) \Delta(q, p_2)
\]
\text{NLEA} specifies not only that the points denoted by \( p_1, p_2 \) are occluded by surfaces, but by the definition of \textit{interior} we also know that it is inconsistent to have

\[
\mathcal{M} \models \text{obscures}(s_2, s_1, t_2)
\]
or

\[
\mathcal{M} \models \text{obscures}(s_5, s_1, t_1)
\]

Note that by \textit{NLEA}, every undepicted point in a surface defines a pose of the surface, since there must exist a depicted surface which obscures the undepicted point. The pose of a surface thus depends on which parts of the surface are occluded and which are depicted. Equivalently, each distinct possible pose of a surface specifies a different set of occluded elements. In the next section, we will specify the clauses corresponding to the depiction and occlusion axioms of \( T_{cw} \). The following notion plays a key role in the definition of those clauses which specify the possible poses of a surface.

\textbf{Definition 9.13} Given an image \( I \), \( \Xi(r) \) is the following set of depiction literals:
\[
\Xi(r) = \{ \Delta(l_1, r) : I \models \text{in}(l_1, r) \text{ and there exists a line } l_2 \in I \text{ such that } I \models (\text{parallel}(l_1, l_2) \lor \text{perp}(l_1, l_2) \lor \text{collinear}(l_1, l_2)) \}.
\]
This is the set of all possible depiction literals for a region in an image: any model of $T_{cw} \cup I$ will satisfy the conjunction of a subset of the literals in $\Xi(r)$ for each region $r$ in the image. For example, in Figure 9.6, $\Xi(r_8)$ is the set

$$\{\lambda(l_{38}, r_8), \lambda(l_{29}, r_8), \lambda(l_{31}, r_8)\}$$

and $\Xi(r_4)$ is the set

$$\{\lambda(l_{34}, r_4), \lambda(l_{23}, r_4)\}$$

$\Xi(r)$ plays a key role in the propositional theory because each subset of $\Xi(r)$ potentially defines a different pose for the surface depicted by the region $r$.

Example: Let $I$ be the image in Figure 9.6. If we have a model $\mathcal{M}_1$ of $T_{cw} \cup I$ such that if we have

$$\mathcal{M}_1 \models \lambda(l_{38}, r_8) \land \lambda(l_{31}, r_8) \land \lambda(r_8, s_8) \land \lambda(r_5, s_5) \land \lambda(r_3, s_3) \land \lambda(r_6, s_6)$$

then we also have

$$\mathcal{M}_1 \models ontop(s_3, s_8) \land ontop(s_5, s_8) \land ontop(s_6, s_8)$$

On the other hand, if we have a model $\mathcal{M}_2$ of $T_{cw} \cup I$ such that

$$\mathcal{M}_2 \models \lambda(l_{38}, r_8) \land \lambda(l_{29}, r_8) \land \lambda(r_8, s_8) \land \lambda(r_1, s_1) \land \lambda(r_2, s_2) \land \lambda(r_4, s_4)$$

then we also have

$$\mathcal{M}_2 \models ontop(s_1, s_8) \land ontop(s_2, s_8) \land ontop(s_4, s_8)$$

Notice that if we have a structure $\mathcal{M}_3$

$$\mathcal{M}_3 \models \lambda(l_{38}, r_8) \land \lambda(l_{29}, r_8) \land \lambda(l_{31}, r_8) \land \lambda(r_8, s_8)$$

then there is no consistent pose for the surface $s_8$, and this structure is not a model of $T_{cw} \cup I$. Thus, there is no model of this image in which all three of the lines $l_{38}, l_{29}$, and $l_{31}$ depict edges in the same surface. □

**Definition 9.14** Given an image $I$ and a region $r$ in the image, suppose $\xi \in \Xi(r)$. A pose for a surface $s$ is consistent with $\xi$ iff the undepicted points in $s$ must have positions included in some region that depicts an occluding surface in the pose, and where these positions are consistent with \textit{PSCA} and the positions of the depicted points in the lines of $\xi$.

It is important to note the role that shape closure plays in the definition of the set of literals in $\Xi(r)$. If two lines are not \textit{parallel} or \textit{perp}, then they cannot depict edges in a rectangular surface, so the only lines that can possibly depict edges in a surface must be either \textit{parallel}, \textit{perp}, or \textit{collinear} \textsuperscript{3}. Also, if we are given a set of lines which can possibly depict edges in the surface, we can then compute the pose using the geometric constraints of \textit{PSCA}.

\textsuperscript{3}Of course, \textit{parallel} or \textit{perp} lines need not depict edges in the same surface, since there may be accidental alignments in the image, but this will be the problem addressed in the next chapter.
9.3.4 The CardWorld Clauses

We are now ready to specify the propositional theory which will be used in the complexity theorems of the next chapter. The language we will use for the clausal theory will use ontop literals (for the occlusion axioms) and \( \Lambda \) literals (for the depiction axioms):

\[
\{ \Delta(r, s), \Lambda(l, r), \text{ontop}(s, s'), \text{abutting}(s, s') \}
\]

Again, since the number of surfaces in the depicted kernel of \( T_{ew} \cup I \) is bounded by \( URDA \), and since we have closure over image elements, there will be a bounded number of depiction and occlusion literals.

**Definition 9.15** Let \( \Pi \) be the set of ground sentences constructed in the following way:

1. For all lines \( l \) such that \( l \models \text{in}(l, r_j) \land \text{in}(l, r_k) \)
   \[
   \{ \Lambda(l, r_j) \lor \Lambda(l, r_k) \} \in \Pi
   \]

   *Clauses of this form will be referred to as figure/ground clauses.*

2. For every two lines \( l_i, l_j \) such that \( l \models \text{in}(l_i, r_m) \land \text{in}(l_j, r_n) \land \neg(\text{collinear}(l_i, l_j) \lor \text{parallel}(l_i, l_j) \lor \text{perp}(l_i, l_j)) \)
   \[
   \neg \Lambda(l_i, r_m) \lor \neg \Lambda(l_j, r_m), \neg \Lambda(l_i, r_n) \lor \neg \Lambda(l_j, r_n) \in \Pi
   \]

3. For all lines \( l_i, l_j \) and regions \( r \) such that \( l \models \text{in}(l_i, r) \land \text{in}(l_j, r) \land \text{collinear}(l_i, l_j) \land (\text{between}(l_i, l_j, r_i) \equiv \neg \text{between}(l_j, l_i, r_j)) \)
   \[
   \neg \Lambda(l_i, r_i) \lor \neg \Lambda(l_j, r_i)
   \]
   \[
   \neg \Lambda(l_j, r_j) \lor \neg \Lambda(l_i, r_j) \in \Pi
   \]

4. For all lines \( l_i, l_j \) and regions \( R \) such that \( l \models \text{in}(l_i, r) \land \theta(a, l_i, l_j) \land a < \pi \land \neg \text{between}(l_i, l_j, r) \)
   \[
   \neg \Lambda(l_i, r) \lor \neg \Lambda(l_j, r) \in \Pi
   \]

5. For every subset \( \xi = \{ \Lambda(l_i, r), ..., \Lambda(l_n, r) \} \) of \( \Xi(r) \), if there is no pose consistent with \( \xi \),
   \[
   \neg \Lambda(l_i, r) \lor ... \lor \neg \Lambda(l_n, r) \in \Pi
   \]

6. For every subset \( \xi = \{ \Lambda(l_i, r), ..., \Lambda(l_n, r) \} \) of \( \Xi(r) \), and if

   \[
   \{ \{ \Delta(r_1, s_1) \land \text{ontop}(s_1, s), ... \}, \{ \Delta(r_n, s_k) \land \text{ontop}(s_n, s), ... \} \}
   \]

   *is the set of poses consistent with \( \xi \), then*

   \[
   \Lambda(l_i, r) \land ... \land \Lambda(l_n, r) \land (\Delta(r_1, s_1) \land \text{ontop}(s_1, s) \land ... \land (\Delta(r_n, s_k) \land \text{ontop}(s_n, s) \land ...)) \lor ... \lor (\Delta(r_n, s_n) \land \text{ontop}(s_n, s) \land \Xi(r) \in \Pi
   \]

7. For every region \( r \) such that \( l \models \text{in}(l_i, r) \land ... \land \text{in}(l_n, r) \)
   \[
   \Delta(r, s) \supset \Lambda(l_i, r) \lor ... \lor \Lambda(l_n, r) \in \Pi
   \]

8. For all surfaces \( s_j, s_k \)
   \[
   \neg \text{ontop}(s_j, s_k) \lor \neg \text{ontop}(s_k, s_j) \in \Pi
   \]
9. For all lines \( L \) and distinct regions \( r_i, r_j \) such that \( I \models \text{in}(L, r_i) \land \text{in}(L, r_j) \),

\[
\Lambda(L, r_i) \lor \neg \Delta(r_i, s_i) \lor \neg \Delta(r_j, s_j) \lor \text{ontop}(s_j, s_i)
\]

\[
\neg \Lambda(L, r_i) \lor \neg \Delta(r_i, s_i) \lor \neg \Delta(r_j, s_j) \lor \neg \text{ontop}(s_j, s_i) \in \Pi
\]

10. For all surface \( s_j, s_k \),

\[
\neg \text{ontop}(s_j, s_k) \lor \neg \text{abutting}(s_k, s_j) \in \Pi
\]

This propositional theory will be useful in designing algorithms for constructing models of the CardWorld axioms. The following theorem shows that these two sets of axioms are indeed logically equivalent.

**Theorem 9.3** Let \( \Phi \) be a sentence in \( \mathcal{L}(\Pi) \). For any planar image \( I \), \( \Pi \cup I \models \Phi \) iff \( T_{cw} \cup RD.A \models \Phi \).

**Proof:** This theorem is a consequence of the following set of lemmas, which demonstrates that those axioms in \( T_{cw} \cup RD.A \) that are equivalent to sentences in \( \mathcal{L}(\Pi) \), are logically equivalent to clauses in \( \Pi \). \( \square \)

**Lemma 9.17**

\[
T_{cw} \models (\forall l, r_1, r_2) (\Lambda(l, r_1) \lor \Lambda(l, r_2)) \equiv (BREA \land NSEA \land ((\forall e, s) \Delta(l, e) \land \text{part}(e, s) \land \text{in}(l, r_1) \land \text{in}(l, r_2) \supset (\Delta(r_1, s) \lor \Delta(r_2, s))))
\]

**Proof:** Recall from the previous chapter that \( BREA \) is equivalent to \( ID.A \) and \( OSD.A \), and that \( ID.A \) and \( NSEA \) are equivalent to the sentence

\[
T_{cw} \models (ID.A \land NSEA) \equiv (\forall l) \text{line}(l) \supset (\exists r) \Lambda(l, r)
\]

By \( OSD.A \) we have

\[
T_{cw} \models (\forall l, e, s) \text{occluding}(e, s) \land \Delta(l, e) \supset (\exists r, r') \text{in}(l, r) \land \text{in}(l, r') \land r \neq r'
\]

By \( ID.A \) we have

\[
T_{cw} I \models (\forall l, e, e', t) \text{abuts}(e, e', t) \land \Delta(l, e) \supset (\exists r, r') \text{in}(l, r) \land \text{in}(l, r') \land r \neq r'
\]

Recall axiom 3.15 of \( T_{image} \), which states that every line is in at most two regions:

\[
(\forall l, r, r', r'' \text{in}(l, r) \land \text{in}(l, r') \land \text{in}(l, r'') \supset r'' = r \lor r'' = r')
\]

By \( ID.A \), we then have

\[
T_{cw} \models (\forall l_1, l_2, r_1, r_2, e_1, e_2, s_1, s_2)(\Delta(l_1, e_1) \land \text{part}(e_1, s_1) \land \Delta(r_1, s_1)) \lor (\Delta(l_1, e_2) \land \text{part}(e_2, s_2) \land \Delta(r_2, s_2))
\]

\[
\equiv \Lambda(l_1, r_1) \lor \Lambda(l_1, r_2)
\]

\( \square \)

Given image closure, it is easy to see that this sentence is equivalent to clauses of the form (1) in the definition of \( \Pi \).

**Lemma 9.18**

\[
T_{cw} \models (\forall l_1, l_2, r_1, r_2)(\neg \Lambda(l_1, r_1) \lor \neg \Lambda(l_2, r_2)) \equiv (RACA \land ((\forall e_1, e_2, a) \theta(a, e_1, e_2) \land \Delta(l_1, e_1) \land \Delta(l_2, e_2) \supset \theta(a, l_1, l_2)))
\]
Proof: By RACA, we need only worry about \textit{collinear}, \textit{parallel}, \textit{perp} lines.

We can combine the depiction axioms for these alignment relations in the following way:

\[ T_{\text{cw}} \cup RDA \models (\forall l, l', r, e, e') \text{ in}(l, r) \land \Delta(l, e) \land \Delta(l', e') \land (e = e' \lor \text{parallel}(e, e') \lor \text{perp}(e, e')) \]

\[ \equiv (\forall l, l', r) \text{ in}(l, r) \land \neg (\text{collinear}(l, l') \lor \text{parallel}(l, l') \lor \text{perp}(l, l')) \lor \neg (\exists e, e'). \Delta(l, e) \land \Delta(l, e') \land (e = e' \lor \text{parallel}(e, e') \lor \text{perp}(e, e')) \]

\[ \equiv (\forall l, l', r) \text{ in}(l, r) \land \text{in}(l', r) \land \neg (\text{collinear}(l, l') \lor \text{parallel}(l, l') \lor \text{perp}(l, l')) \lor \neg (\exists e, e', s) \text{ part}(e, s) \land \text{part}(e', s) \land \Delta(r, s) \land \Delta(l, e) \land \Delta(l', e')) \]

since by \textit{URDA}, \textit{r} must depict only the surface containing the edges \( e, e' \). Therefore, by the definition of \( \Lambda \), we have

\[ T_{\text{cw}} \cup RDA \models (\forall l, l', r) \text{ in}(l, r) \land \text{in}(l', r) \land \neg (\text{collinear}(l, l') \lor \text{parallel}(l, l') \lor \text{perp}(l, l')) \lor \neg (\exists e, e', s) \text{ part}(e, s) \land \text{part}(e', s) \land \Delta(r, s) \land \Delta(l, e) \land \Delta(l', e')) \]

\[ \equiv \neg (\Lambda(l, r) \land \Lambda(l', r)) \]

\[ \square \]

Given image closure, it is easy to see that this sentence is equivalent to clauses of the form (2) in the definition of \( \Pi \).

Lemma 9.19

\[ T_{\text{cw}} \models (\forall l, l_2, r_1, r_2) (((\neg \Lambda(l_1, r_1) \lor \neg \Lambda(l_2, r_2)) \land (\neg \Lambda(l_1, r_2) \lor \neg \Lambda(l_2, r_1))) \]

\[ \equiv (\forall e, s) \text{ in}(l_1, r_1) \land \text{in}(l_2, r_2) \land \Delta(l_1, e) \land \Delta(l_2, e) \land \text{part}(e, s) \lor \]

\[ [(\Delta(r_1, s) \equiv \Delta(r_2, s)] \equiv (\text{between}(l_1, l_2, r_1) \equiv \neg \text{between}(l_2, l_1, r_2))] \]

Proof: We have

\[ T_{\text{cw}} \models (\forall l, l', r, r', e, s) \text{ in}(l, r) \land \text{in}(l', r') \land \Delta(l, e) \land \Delta(l', e') \land \text{part}(e, s) \land e \neq e' \lor \]

\[ [(\Delta(r, s) \equiv \Delta(r', s)] \equiv (\text{between}(l, l', r) \equiv \neg \text{between}(l', l, r')) ] \]

\[ \equiv (\forall l, l', r, r', e, s) \text{ in}(l, r) \land \text{in}(l', r') \land \text{between}(l, l', r) \equiv \text{between}(l', l, r') ) \lor \]

\[ (\Delta(l, e) \land \Delta(l', e') \land \text{part}(e, s) \lor (\Delta(r, s) \equiv \neg \Delta(r', s))] \]

\[ \equiv (\forall l, l', r, r') (\Lambda(l, r) \lor \neg \Lambda(l', r)) \land (\Lambda(l', r') \lor \neg \Lambda(l, r')) \]

\[ \square \]

Given image closure, it is easy to see that this sentence is equivalent to clauses of the form (3) in the definition of \( \Pi \).

Lemma 9.20

\[ T_{\text{cw}} \models (\forall l, l_2, r) \neg \Lambda(l_1, r) \lor \neg \Lambda(l_2, r)) \equiv \]

\[ [(\forall e, s, l, l', r, a) \text{ part}(e, s) \land \text{part}(e', s) \land \Delta(l, e) \land \Delta(l', e') \land \text{in}(l, r) \land \theta(a, l, l') \lor \]

\[ ((\Delta(r, s) \equiv \text{between}(l, l', r)) \equiv a < \pi]) \]
Proof:

\[ T_{cw} \models \\
(\forall l', l, r, e, e', s) \Delta(l, e) \land \Delta(l', e') \land \text{in}(l, r) \land \text{part}(e, s) \land \text{part}(e', s) \land \theta(a, l, l') \]

\[ \lor (\Delta(r, s) \equiv \text{between}(l_1, l_2, r) \land a < \pi) \]

\[ \equiv (\forall l', l, r) \neg \text{between}(l, l', r) \land \text{in}(l, r) \lor (\exists e, e', s) \text{part}(e, s) \land \text{part}(e', s) \land \Delta(l, e) \land \Delta(l', e') \land \Delta(r, s) \]

Using the definition of \( \Lambda \), we can rewrite this as:

\[ T_{cw} \models (\forall l_1, l_2, r, s) \text{in}(l_1, r) \land \neg \text{between}(l_1, l_2, r) \lor \neg (\Lambda(l_1, r) \land \Lambda(l_2, r)) \]

\( \Box \)

Given image closure, it is easy to see that this sentence is equivalent to clauses of the form (4) in the definition of \( \Pi \).

Lemma 9.21  \( \text{RD}A \) is equivalent to clauses of the form (7) in the definition of \( \Pi \).

Proof: This follows from image closure and the definition of \( \text{RD}A \) which is equivalent to

\[ (\forall r, s) \text{surface}(s) \land \Delta(r, s) \lor (\exists l, e) \Delta(l, e) \land \text{part}(e, s) \land \text{in}(l, r) \land (l = l_1 \lor \ldots \lor l = l_n) \]

where \( l_1, \ldots, l_n \) are constants denoting lines in the image \( I \). \( \Box \)

The next lemma shows that \( \text{DTR} \) is equivalent to clauses of the form (9) in the definition of \( \Pi \).

Lemma 9.22

\[ T_{cw} \models (\forall l, r, s_1, s_2) ((\Lambda(l, r) \lor \text{ontop}(s_2, s_1)) \land (\neg \Lambda(l, r) \lor \neg \text{ontop}(s_2, s_1))) \equiv \text{DTR} \]

Proof: Recall that \( \text{DTR} \) is the sentence

\[ (\forall l, e, r, s) \text{in}(l, r) \land \Delta(l, e) \land \Delta(r, s) \lor \text{occluding}(e, s) \lor \text{part}(e, s) \lor (\exists e', t) \text{abuts}(e, e', t) \]

By \( \text{CS}A \) and the definition of \( \Lambda \), we have

\[ T_{cw} \cup I \models [(\forall l, e, r_1, r_2, s_1, s_2) \text{in}(l, r_1) \land \text{in}(l, r_2) \land \Delta(l, e) \land \Delta(r_1, s_1) \land \Delta(r_2, s_2) \land \text{part}(e, s) \lor \text{occluding}(e, s) \lor \text{abutting}(s_1, s_2)] \]

\[ \equiv (\forall l, r_1, r_2, s_1, s_2) \Lambda(l, r_1) \land \Delta(r_1, s_1) \land \Delta(r_2, s_2) \lor \neg \text{ontop}(s_2, s_1) \]

and

\[ T_{cw} \cup I \models [(\forall l, e, r_1, r_2, s_1, s_2) \text{in}(l, r_1) \land \text{in}(l, r_2) \land \Delta(l, e) \land \Delta(r_1, s_1) \land \Delta(r_2, s_2) \land \neg \text{part}(e, s) \lor \text{occluding}(e, s)] \]

\[ \equiv (\forall l, r_1, r_2, s_1, s_2) \neg \Lambda(l, r_1) \land \Delta(r_1, s_1) \land \Delta(r_2, s_2) \lor \text{ontop}(s_2, s_1) \]

\( \Box \)

Given image closure, it is easy to see that this sentence is equivalent to clauses of the form (9) in the definition of \( \Pi \).

We can now use the notion of the pose of a surface to define a set of clauses corresponding to the occlusion axioms:
Lemma 9.23 Given a planar image $I$, the sentence
\[(\forall r, s) \neg (\exists t \text{ interior}(s, t) \land \text{ inclusion}(r, t)) \supset \neg \Delta(r, s)\]
is logically equivalent to the conjunction of clauses of the form (5) in the definition of $\Pi$.

Proof: Let $\mathcal{M}$ be a model of $T_{cw} \cup \text{ RDA } \cup I$ for some planar image $I$. For any variable assignment $\sigma$ we have
\[\mathcal{M}, \sigma \models \neg (\exists t \text{ interior}(s, t) \land \text{ inclusion}(r, t))\]
iff the surface $\sigma(s)$ does not have a consistent pose in $\mathcal{M}$ for any subset of $\Xi(\sigma(r))$.
By condition (5) in the construction of $\Pi$, we must therefore have
\[\mathcal{M}, \sigma \models \neg \Lambda(l_1, r) \lor \ldots \lor \neg \Lambda(l_i, r)\]
for every subset of $\Xi(\sigma(r))$, which is equivalent to
\[\mathcal{M}, \sigma \models \neg \Delta(r, s)\]
by $\text{ RDA}$ and the definition of $\Lambda$. $\square$

Lemma 9.24 Given a planar image $I$, the sentence
\[(\forall r, s) ((\exists t \text{ interior}(s, t) \land \text{ inclusion}(r, t)) \land \neg \Delta(r, s))\]
\[\supset (\exists s') \Delta(r, s') \land \text{ onttop}(s', s)\]
is logically equivalent to the conjunction of clauses of the form (6) in the definition of $\Pi$.

Proof: Let $\mathcal{M}$ be a model of $T_{cw} \cup \text{ RDA } \cup I$ for some planar image $I$. For any variable assignment $\sigma$ we have
\[\mathcal{M}, \sigma \models (\exists t) \text{ interior}(s, t) \land \text{ inclusion}(r, t) \land \neg \Delta(r, s)\]
iff there is a consistent pose for $\sigma(s)$ given some subset of $\Xi(\sigma(r))$.
By condition (6) in the construction of $\Pi$, we must therefore have
\[\mathcal{M}, \sigma \models \Lambda(l_1, r) \land \ldots \land \Lambda(l_n, r) \supset (\Delta(r_k, s_k) \land \text{ onttop}(s_k, s) \land \ldots) \lor \ldots \lor (\Delta(r_n, s_n) \land \text{ onttop}(s_n, s))\]
for some subset of $\Xi(\sigma(r))$, which is equivalent to
\[\mathcal{M}, \sigma \models (\exists s') \Delta(r, s') \land \text{ onttop}(s', s)\]

$\square$

Also note that there are only a linear number of pose clauses in $\Pi$:

Lemma 9.25 Given an assignment of surface depiction literals to all regions in an image $I$, there are at most $4|l|$ pose clauses for each surface, where $|l|$ is the number of lines in $I$.

Proof: Since there are only four edges in a rectangle, there can be at most four sets of non-collinear lines that depict edges in a surface depicted by some set of regions. Thus there are at most $2^4$ sets of depicted edges. There are $|l|/4$ sets of parallel and perp lines in these regions, so that we have
\[2^4 \times |l|/4 = 4|l|\]
pose clauses. $\square$

We will now illustrate the above transformation from first-order theory to a set of propositional clauses.
Figure 9.6: Image for the example of a clausal CardWorld theory.
Example: Consider the image in Figure 9.6. For this image \( I, T_{cw} \cup I \cup RD.A \) is equivalent to the following set of clauses:

Depiction clauses:

\[
\begin{align*}
\Lambda(l_1, r_1), \Lambda(l_2, r_1), \Lambda(l_3, r_2), \Lambda(l_4, r_3), \Lambda(l_5, r_3), \Lambda(l_6, r_3), \Lambda(l_7, r_3), \Lambda(l_8, r_3) \\
\Lambda(l_9, r_6), \Lambda(l_{10}, r_6), \Lambda(l_{11}, r_7), \Lambda(l_{12}, r_7), \Lambda(l_{13}, r_{10}), \Lambda(l_{14}, r_{10}), \Lambda(l_{15}, r_{10}) \\
\Lambda(l_{16}, r_1), \Lambda(l_{17}, r_2), \neg \Lambda(l_{17}, r_1), \Lambda(l_{18}, r_2), \neg \Lambda(l_{18}, r_1), \Lambda(l_{19}, r_2), \neg \Lambda(l_{19}, r_1) \\
\Lambda(l_{21}, r_2), \neg \Lambda(l_{21}, r_3), \Lambda(l_{22}, r_2), \neg \Lambda(l_{22}, r_3), \Lambda(l_{23}, r_8), \neg \Lambda(l_{23}, r_3), \\
\Lambda(l_{27}, r_6), \neg \Lambda(l_{27}, r_3), \Lambda(l_{35}, r_{10}), \neg \Lambda(l_{35}, r_1), \Lambda(l_{36}, r_{10}), \neg \Lambda(l_{36}, r_1) \\
\Lambda(l_{20}, r_2) \vee \Lambda(l_{20}, r_4), \Lambda(l_{23}, r_3) \vee \Lambda(l_{23}, r_4) \\
\Lambda(l_{25}, r_3) \vee \Lambda(l_{24}, r_5), \Lambda(l_{28}, r_5) \vee \Lambda(l_{29}, r_6) \\
\Lambda(l_{29}, r_5) \vee \Lambda(l_{29}, r_6), \Lambda(l_{30}, r_6) \vee \Lambda(l_{30}, r_6) \\
\Lambda(l_{31}, r_4) \vee \Lambda(l_{31}, r_8), \Lambda(l_{32}, r_4) \vee \Lambda(l_{32}, r_10) \\
\Lambda(l_{34}, r_1) \vee \Lambda(l_{34}, r_4), \Lambda(l_{37}, r_9) \vee \Lambda(l_{37}, r_10) \\
\Lambda(l_{38}, r_9) \vee \Lambda(l_{38}, r_9), \Lambda(l_{39}, r_7) \vee \Lambda(l_{39}, r_9) \\
\Lambda(l_{40}, r_7) \vee \Lambda(l_{40}, r_8), \Lambda(l_{41}, r_7) \vee \Lambda(l_{41}, r_{10}) \\
\neg \Lambda(l_{30}, r_4) \vee \neg \Lambda(l_{23}, r_4), \neg \Lambda(l_{20}, r_4) \vee \neg \Lambda(l_{23}, r_4) \\
\neg \Lambda(l_{20}, r_4) \vee \neg \Lambda(l_{31}, r_4), \neg \Lambda(l_{20}, r_4) \vee \neg \Lambda(l_{32}, r_4) \\
\neg \Lambda(l_{20}, r_4) \vee \neg \Lambda(l_{34}, r_4), \neg \Lambda(l_{23}, r_4) \vee \neg \Lambda(l_{31}, r_4) \\
\neg \Lambda(l_{23}, r_4) \vee \neg \Lambda(l_{32}, r_4) \\
\neg \Lambda(l_{31}, r_4) \vee \neg \Lambda(l_{32}, r_4), \neg \Lambda(l_{31}, r_4) \vee \neg \Lambda(l_{34}, r_4) \\
\neg \Lambda(l_{32}, r_4) \vee \neg \Lambda(l_{34}, r_4), \neg \Lambda(l_{25}, r_5) \vee \neg \Lambda(l_{28}, r_5) \\
\neg \Lambda(l_{25}, r_5) \vee \neg \Lambda(l_{29}, r_5), \neg \Lambda(l_{28}, r_5) \vee \neg \Lambda(l_{29}, r_5) \\
\neg \Lambda(l_{37}, r_9) \vee \neg \Lambda(l_{38}, r_9), \neg \Lambda(l_{37}, r_9) \vee \neg \Lambda(l_{39}, r_9) \\
\neg \Lambda(l_{38}, r_9) \vee \neg \Lambda(l_{39}, r_9)
\end{align*}
\]

Clauses arising from inconsistent poses:

\[
\begin{align*}
\neg \Lambda(l_{23}, r_4) \vee \neg \Lambda(l_{34}, r_4) \\
\neg \Lambda(l_{31}, r_9) \vee \neg \Lambda(l_{38}, r_8) \vee \neg \Lambda(l_{29}, r_8)
\end{align*}
\]
\neg \Lambda(l_{38}, r_9)

Pose clauses for surfaces with consistent poses (to enhance readability, we will use propositional letters \( \rho_i \) to denote poses):

\[ \Lambda(l_{20}, r_4) \supset \rho_1 \lor \rho_2 \]

\[ \rho_1 \equiv \Delta(r_1, s_1) \land ontop(s_1, s_4) \land \Delta(r_3, s_3) \land ontop(s_3, s_4) \land \Delta(r_4, s_4) \land ontop(s_4, s_4) \]

\[ \land \Delta(r_5, s_5) \land ontop(s_5, s_4) \land \Delta(r_7, s_7) \land ontop(s_7, s_4) \land \Delta(r_8, s_8) \land ontop(s_8, s_4) \land \Delta(r_9, s_9) \land ontop(s_9, s_4) \land \Delta(r_{10}, s_{10}) \land ontop(s_{10}, s_4) \]

\[ \rho_2 \equiv \Delta(r_1, s_1) \land ontop(s_1, s_4) \land \Delta(r_3, s_3) \land ontop(s_3, s_4) \land \Delta(r_4, s_4) \land ontop(s_4, s_4) \]

\[ \land \Delta(r_5, s_5) \land ontop(s_5, s_4) \land \Delta(r_7, s_7) \land ontop(s_7, s_4) \land \Delta(r_8, s_8) \land ontop(s_8, s_4) \land \Delta(r_9, s_9) \land ontop(s_9, s_4) \land \Delta(r_{10}, s_{10}) \land ontop(s_{10}, s_4) \]

\[ \Lambda(l_{34}, r_4) \supset \rho_3 \lor \rho_4 \lor \rho_5 \]

\[ \rho_3 \equiv \Delta(r_2, s_2) \land ontop(s_2, s_4) \land \Delta(r_8, s_8) \land ontop(s_8, s_4) \land \Delta(r_9, s_9) \land ontop(s_9, s_4) \]

\[ \rho_4 \equiv \Delta(r_2, s_2) \land ontop(s_2, s_4) \land \Delta(r_8, s_8) \land ontop(s_8, s_4) \land \Delta(r_9, s_9) \land ontop(s_9, s_4) \land \Delta(r_{10}, s_{10}) \land ontop(s_{10}, s_4) \]

\[ \rho_5 \equiv \Delta(r_2, s_2) \land ontop(s_2, s_4) \land \Delta(r_3, s_3) \land ontop(s_3, s_4) \land \Delta(r_5, s_5) \land ontop(s_5, s_4) \land \Delta(r_6, s_6) \land ontop(s_6, s_4) \]

\[ \land \Delta(r_8, s_8) \land ontop(s_8, s_4) \land \Delta(r_{10}, s_{10}) \land ontop(s_{10}, s_4) \]

\[ \Lambda(l_{33}, r_4) \supset \rho_6 \lor \rho_7 \lor \rho_8 \]

\[ \rho_6 \equiv \Delta(r_2, s_2) \land ontop(s_2, s_4) \land \Delta(r_8, s_8) \land ontop(s_8, s_4) \land \Delta(r_9, s_9) \land ontop(s_9, s_4) \]

\[ \rho_7 \equiv \Delta(r_2, s_2) \land ontop(s_2, s_4) \land \Delta(r_8, s_8) \land ontop(s_8, s_4) \land \Delta(r_9, s_9) \land ontop(s_9, s_4) \land \Delta(r_{10}, s_{10}) \land ontop(s_{10}, s_4) \]

\[ \rho_8 \equiv \Delta(r_2, s_2) \land ontop(s_2, s_4) \land \Delta(r_1, s_1) \land ontop(s_1, s_4) \land \Delta(r_9, s_9) \land ontop(s_9, s_4) \land \Delta(r_{10}, s_{10}) \land ontop(s_{10}, s_4) \]

\[ \Lambda(l_{31}, r_4) \supset \rho_9 \]

\[ \rho_9 \equiv \Delta(r_1, s_1) \land ontop(s_1, s_4) \land \Delta(r_2, s_2) \land ontop(s_2, s_4) \land \Delta(r_3, s_3) \land ontop(s_3, s_4) \land \Delta(r_{10}, s_{10}) \land ontop(s_{10}, s_4) \]

\[ \Lambda(l_{32}, r_4) \supset \rho_{10} \]

\[ \rho_{10} \equiv \Delta(r_1, s_1) \land ontop(s_1, s_4) \land \Delta(r_2, s_2) \land ontop(s_2, s_4) \land \Delta(r_3, s_3) \land ontop(s_3, s_4) \]

\[ \land \Delta(r_5, s_5) \land ontop(s_5, s_4) \land \Delta(r_8, s_8) \land ontop(s_8, s_4) \]

\[ \Lambda(l_{29}, r_5) \supset \rho_{11} \]
\[ \rho_{11} \equiv \Delta(r_4, s_4) \land \text{ontop}(s_4, s_5) \land \Delta(r_2, s_2) \land \text{ontop}(s_2, s_5) \land \Delta(r_3, s_3) \land \text{ontop}(s_3, s_5) \land \Delta(r_6, s_6) \land \text{ontop}(s_6, s_6) \]

\[ \Lambda(l_{25}, r_5) \supset \rho_{12} \]

\[ \rho_{12} \equiv \Delta(r_4, s_4) \land \text{ontop}(s_4, s_5) \land \Delta(r_6, s_6) \land \text{ontop}(s_6, s_5) \land \Delta(r_8, s_8) \land \text{ontop}(s_8, s_5) \land \Delta(r_6, s_6) \land \text{ontop}(s_6, s_6) \]

\[ \Lambda(l_{28}, r_5) \supset \rho_{13} \]

\[ \rho_{13} \equiv \Delta(r_4, s_4) \land \text{ontop}(s_4, s_5) \land \Delta(r_3, s_3) \land \text{ontop}(s_3, s_5) \land \Delta(r_8, s_8) \land \text{ontop}(s_8, s_5) \]

\[ \Lambda(l_{39}, r_9) \supset \rho_{14} \land \rho_{15} \]

\[ \rho_{14} \equiv \Delta(r_1, s_1) \land \text{ontop}(s_1, s_9) \land \Delta(r_4, s_4) \land \text{ontop}(s_4, s_9) \land \Delta(r_5, s_5) \land \text{ontop}(s_5, s_9) \land \Delta(r_8, s_8) \land \text{ontop}(s_8, s_9) \land \Delta(r_{10}, s_{10}) \land \text{ontop}(s_{10}, s_9) \]

\[ \rho_{15} \equiv \Delta(r_1, s_1) \land \text{ontop}(s_1, s_9) \land \Delta(r_4, s_4) \land \text{ontop}(s_4, s_9) \land \Delta(r_2, s_2) \land \text{ontop}(s_2, s_9) \land \Delta(r_8, s_8) \land \text{ontop}(s_8, s_9) \land \Delta(r_{10}, s_{10}) \land \text{ontop}(s_{10}, s_9) \]

\[ \Lambda(l_{37}, r_9) \supset \rho_{16} \]

\[ \rho_{16} \equiv \Delta(r_3, s_3) \land \text{ontop}(s_3, s_9) \land \Delta(r_4, s_4) \land \text{ontop}(s_4, s_9) \land \Delta(r_5, s_5) \land \text{ontop}(s_5, s_9) \land \Delta(r_8, s_8) \land \text{ontop}(s_8, s_9) \land \Delta(r_7, s_7) \land \text{ontop}(s_7, s_9) \]

\[ \Lambda(l_{31}, r_8) \land \Lambda(l_{38}, r_8) \supset \rho_{17} \]

\[ \rho_{17} \equiv \Delta(r_3, s_3) \land \text{ontop}(s_3, s_8) \land \Delta(r_4, s_4) \land \text{ontop}(s_4, s_8) \land \Delta(r_5, s_5) \land \text{ontop}(s_5, s_8) \land \Delta(r_6, s_6) \land \text{ontop}(s_6, s_8) \land \Delta(r_7, s_7) \land \text{ontop}(s_7, s_8) \]

\[ \Lambda(l_{31}, r_8) \land \Lambda(l_{29}, r_8) \supset \rho_{18} \]

\[ \rho_{18} \equiv \Delta(r_3, s_3) \land \text{ontop}(s_3, s_8) \land \Delta(r_9, s_9) \land \text{ontop}(s_9, s_8) \land \Delta(r_5, s_5) \land \text{ontop}(s_5, s_8) \land \Delta(r_6, s_6) \land \text{ontop}(s_6, s_8) \land \Delta(r_7, s_7) \land \text{ontop}(s_7, s_8) \land \Delta(r_{10}, s_{10}) \land \text{ontop}(s_{10}, s_8) \]

\[ \Lambda(l_{38}, r_8) \land \Lambda(l_{29}, r_8) \supset \rho_{19} \]

\[ \rho_{19} \equiv \Delta(r_1, s_1) \land \text{ontop}(s_1, s_8) \land \Delta(r_9, s_9) \land \text{ontop}(s_9, s_8) \land \Delta(r_4, s_4) \land \text{ontop}(s_4, s_8) \land \Delta(r_6, s_6) \land \text{ontop}(s_6, s_8) \land \Delta(r_7, s_7) \land \text{ontop}(s_7, s_8) \]

\[ \square \]
9.4 Satisfying the Pose Clauses

We have observed that there are two basic classes of clauses within a propositional CardWorld theory $\Pi$ - the depiction literals $\Lambda(l, r)$ and the occlusion literals $\text{ontop}(s_1, s_2)$. In this section, we find that if we have already found a satisfying interpretation for the depiction clauses, then there is a polynomial algorithm for satisfying the pose clauses (which contain the $\text{ontop}$ literals).

Any model satisfying the depiction clauses consists of an assignment of edge depiction literals to the lines in an image $I$ such that for every line $l$ such that $I \models \text{in}(l, r_1) \land \text{in}(l, r_2)$ there are two edge depiction literals - one of $\Lambda(l, r_1)$ or $\neg \Lambda(l, r_1)$, and one of $\Lambda(l, r_2)$ or $\neg \Lambda(l, r_2)$. This assignment also provides an assignment of surface depiction literals to the regions of $I$ in the following way: if $\Lambda(l, r_1)$ then $\Delta(r_1, s_1)$: if for every line $l$ such that $I \models \text{in}(l, r)$ we have $\neg \Lambda(l, r)$, we have $(\forall s) \neg \Delta(r, s)$ by RDA.

**Definition 9.16** A complete assignment of edge depiction literals to the lines in an image $I$ is a set of edge depiction literals $\Sigma$ such that for all lines $L$ and regions $r, r'$, if

$$I \models \text{in}(l, r) \land \text{in}(l, r')$$

then we have

$$\Pi \cup \Sigma \models \Lambda(l, r) \land \Lambda(l, r')$$

or

$$\Pi \cup \Sigma \models \Lambda(l, r) \land \neg \Lambda(l, r')$$

or

$$\Pi \cup \Sigma \models \neg \Lambda(l, r) \land \Lambda(l, r')$$

**Definition 9.17** A complete assignment of surface depiction literals to the regions in an image $I$ is a set of surface depiction literals $\Sigma$ such that for all regions $r$ either there exists a surface $s$ such that

$$\Pi \cup \Sigma \models \Delta(r, s)$$

or

$$\Pi \cup \Sigma \models \neg (\exists s') \Delta(r, s')$$

These assignments of depiction literals will satisfy the depiction clauses; the remaining work to find a model consists in satisfying the occlusion clauses. This section will study the complexity of this latter problem, given that the depiction axioms have already been satisfied and depiction literals have been completely assigned. The next chapter will study the complexity of finding complete assignments of depiction literals.

We will decompose the problem of satisfying the occlusion clauses into two tasks. The first, which we will call SELECT-POSE, finds a consistent pose for a surface if one exists. The second task, which we will call POSE CONSISTENCY, reassigns the surface depiction literals if the given assignment does not lead to a consistent pose for a surface.

**Definition 9.18** **SELECT-POSE TASK**

Given a complete assignment of surface depiction literals to the regions in $I$ and the pose clauses for the regions depicting surfaces, find a pose for each surface consistent with respect to $T \cup I \cup \text{RDA}$ or identify the surfaces with no consistent poses.
**Definition 9.19** **POSE CONSISTENCY TASK:**
Given a complete assignment of surface depiction literals to all regions in $I$ and a complete assignment of edge depiction literals to all lines in $I$, find a pose for every surface consistent with respect to $T \cup f \cup RDA$.

We will first consider the SELECT-POSE task. Since all depiction literals have been completely assigned, the only clauses left in $\Pi$ are those containing $\text{ontop}$ literals. Further, there are only two kinds of clauses in $\Pi$ containing $\text{ontop}$ literals – the positive disjunctive pose clauses, and the binary negative clauses which are equivalent to the sentence

$$\forall s_1, s_2 \text{ ontop}(s_1, s_2) \supset \neg \text{ontop}(s_2, s_1)$$

Thus, we can represent the remaining clauses in $\Pi$ as a graph:

**Definition 9.20** Given an image $I$, let $P$ be the set of pose clauses in $\Pi$. An occlusion graph is a graph $G = (V, E)$ such that $V = \{p_i | p_i$ is a pose in a pose clause in $P\}$ and $E = \{(p_1, p_2) | \neg p_1 \lor \neg p_2 \text{ is a clause in } \Pi\}$.

**Example:** Consider the examples in Figure 9.7. In (a), we have the pose clauses

$$\Delta(r_1, s_1) \supset p_1$$
$$\Delta(r_2, s_2) \supset p_2$$

where

$$p_1 \equiv \Delta(r_2, s_2) \land \text{ontop}(s_2, s_1)$$
$$p_2 \equiv \Delta(r_1, s_1) \land \text{ontop}(s_1, s_2)$$

Notice that we have $p_1 \supset \neg p_2$, since the surfaces cannot be mutually occluding. The occlusion graph for these clauses is given below the image. Image (b) has the same set of pose clauses and occlusion graph. The circles around the pose literals $p_i$ indicate the set of possible poses for a particular surface.

In Figure 9.7(c) we have the pose clauses

$$\Delta(r_1, s_1) \supset p_2 \lor p_3$$
$$\Delta(r_2, s_2) \supset p_1$$
$$\Delta(r_3, s_3) \supset p_4$$

where

$$p_1 \equiv \Delta(r_1, s_1) \land \text{ontop}(s_1, s_2)$$
$$p_2 \equiv \Delta(r_2, s_2) \land \text{ontop}(s_2, s_1)$$
$$p_3 \equiv \Delta(r_3, s_3) \land \text{ontop}(s_3, s_1)$$
Figure 9.7: Examples of occlusion graphs.
Figure 9.8: Additional examples of occlusion graphs.
\[ \rho_4 \equiv \Delta(r_1, s_1) \land \text{ontop}(s_1, s_2) \]

Again, because the surfaces cannot be mutually occluding, we have the following:

\[ \neg \rho_1 \lor \neg \rho_2 \lor \neg \rho_3 \lor \neg \rho_4 \]

Also, every surface has a unique pose, so that if \( s_1 \) has the pose \( \rho_3 \) in some model, then this model cannot also satisfy the pose \( \rho_2 \). The occlusion graph for these clauses is given below the image.

In Figure 9.7(d) we have the pose clauses

\[ \Delta(r_1, s_1) \supset \Delta(r_2, s_2) \land \text{ontop}(s_2, s_1) \]

\[ \Delta(r_2, s_2) \supset \]

\[ \Delta(r_1, s_1) \land \text{ontop}(s_1, s_2) \lor \Delta(r_3, s_3) \land \text{ontop}(s_3, s_2) \lor \Delta(r_1, s_1) \land \Delta(r_3, s_3) \land \text{ontop}(s_1, s_2) \land \text{ontop}(s_3, s_2) \]

\[ \Delta(r_3, s_3) \supset \Delta(r_2, s_2) \land \text{ontop}(s_2, s_3) \]

where

\[ \rho_1 \equiv \Delta(r_2, s_2) \land \text{ontop}(s_2, s_1) \]

\[ \rho_2 \equiv \Delta(r_1, s_1) \land \text{ontop}(s_1, s_2) \]

\[ \rho_3 \equiv \Delta(r_3, s_3) \land \text{ontop}(s_3, s_2) \]

\[ \rho_4 \equiv \Delta(r_1, s_1) \land \Delta(r_3, s_3) \land \text{ontop}(s_1, s_2) \land \text{ontop}(s_3, s_2) \]

\[ \rho_5 \equiv \Delta(r_2, s_2) \land \text{ontop}(s_2, s_3) \]

Because the surfaces cannot be mutually occluding, we have

\[ \neg \rho_1 \lor \neg \rho_2 \lor \neg \rho_3 \lor \neg \rho_5 \lor \neg \rho_3 \lor \neg \rho_4 \]

Since the three possible poses for the surface \( s_2 \) are mutually inconsistent, we have

\[ \neg \rho_2 \lor \neg \rho_3 \lor \neg \rho_4 \lor \neg \rho_3 \lor \neg \rho_4 \]

The occlusion graph for these clauses is given below the image.

In Figure 9.8(a) we have the pose clauses

\[ \Delta(r_1, s_1) \supset \rho_1 \]

\[ \Delta(r_2, s_2) \supset \rho_2 \lor \rho_3 \lor \rho_4 \]

\[ \Delta(r_3, s_3) \supset \rho_5 \]

where

\[ \rho_1 \equiv \Delta(r_2, s_2) \land \text{ontop}(s_2, s_1) \land \Delta(r_3, s_3) \land \text{ontop}(s_3, s_1) \]
\[ \rho_2 \equiv \Delta (r_1, s_1) \land \text{ontop}(s_1, s_2) \]
\[ \rho_3 \equiv \Delta (r_2, s_3) \land \text{ontop}(s_3, s_2) \]
\[ \rho_4 \equiv \Delta (r_1, s_1) \land \Delta (r_2, s_3) \land \text{ontop}(s_1, s_2) \land \text{ontop}(s_3, s_2) \]
\[ \rho_5 \equiv \Delta (r_2, s_2) \land \text{ontop}(s_2, s_3) \land \Delta (r_1, s_1) \land \text{ontop}(s_1, s_3) \]

Because the surfaces cannot be mutually occluding, we have
\[ \neg \rho_1 \lor \neg \rho_2 \lor \neg \rho_3 \lor \neg \rho_4 \lor \neg \rho_5 \lor \neg \rho_6 \]

Since the three possible poses for the surface \( s_2 \) are mutually inconsistent, we have
\[ \neg \rho_2 \lor \neg \rho_3 \lor \neg \rho_4 \lor \neg \rho_5 \lor \neg \rho_6 \]

The occlusion graph for these clauses is given below the image.

In Figure 9.8(b) we have the pose clauses
\[ \Delta (r_1, s_1) \supset \Delta (r_2, s_2) \land \text{ontop}(s_2, s_1) \lor \Delta (r_3, s_3) \land \text{ontop}(s_3, s_1) \]
\[ \Delta (r_2, s_2) \supset \Delta (r_1, s_1) \land \text{ontop}(s_1, s_2) \lor \Delta (r_4, s_4) \land \text{ontop}(s_4, s_2) \]
\[ \Delta (r_3, s_3) \supset \Delta (r_1, s_1) \land \text{ontop}(s_1, s_3) \]
\[ \Delta (r_4, s_4) \supset \Delta (r_2, s_2) \land \text{ontop}(s_2, s_4) \]

where
\[ \rho_1 \equiv \Delta (r_2, s_2) \land \text{ontop}(s_2, s_1) \]
\[ \rho_2 \equiv \Delta (r_3, s_3) \land \text{ontop}(s_3, s_1) \]
\[ \rho_3 \equiv \Delta (r_1, s_1) \land \text{ontop}(s_1, s_2) \]
\[ \rho_4 \equiv \Delta (r_4, s_4) \land \text{ontop}(s_4, s_2) \]
\[ \rho_5 \equiv \Delta (r_1, s_1) \land \text{ontop}(s_1, s_3) \]
\[ \rho_6 \equiv \Delta (r_2, s_2) \land \text{ontop}(s_2, s_4) \]

Since the surfaces cannot be mutually occluding, we have
\[ \neg \rho_1 \lor \neg \rho_2 \lor \neg \rho_3 \lor \neg \rho_4 \lor \neg \rho_5 \lor \neg \rho_6 \]

Since the two possible poses for the surface \( s_1 \) are mutually inconsistent, we have
\[ \neg \rho_1 \lor \neg \rho_2 \]

and since the two possible poses for the surface \( s_2 \) are mutually inconsistent, we have
\[ \neg \rho_4 \lor \neg \rho_5 \]

The occlusion graph for these clauses is given below the image. □
Example: We can also consider the image in Figure 9.6: the pose clauses for this image were given in the previous section. We can consider various complete assignments of edge depiction literals to construct different occlusion graphs. In Figure 9.8(c), we consider the assignment \( \Lambda(l_{29}, r_4), \Lambda(l_{39}, r_9) \). In (d), we consider the assignment \( \Lambda(l_{34}, r_4), \Lambda(l_{39}, r_9) \). In (e), we consider the assignment \( \Lambda(l_{23}, r_4), \Lambda(l_{39}, r_9) \). In (f), we consider the assignment \( \Lambda(l_{31}, r_4), \Lambda(l_{39}, r_9) \). □

The intuitions of these examples can be formalized by the following lemma:

Lemma 9.26 Let \( G = (V, E) \) be an occlusion graph. Then \( V' \subset V \) is an independent set of size \( |\Sigma| \) iff every surface has a consistent pose, where \( |\Sigma| \) is the number of surfaces.

Proof: Note that all occlusion clauses are either positive pose clauses or negative binary clauses corresponding to mutually inconsistent poses of two surfaces. By the construction of \( \Pi \), these are the only occlusion clauses.

Every pose appears in a unique pose clause, since each pose is the set of occluding surfaces for a given surface, and the pose clause contains all possible poses for a surface. Thus every pose corresponds to a unique vertex in the occlusion graph.

All poses in the negative clauses appear in pose clauses, by construction of \( \Pi \). Thus no new vertices in the occlusion graph are introduced by the negative clauses.

Poses in the pose clause for a given surface are mutually exclusive, since each pose is the set of occluding surfaces for different sets of undepicted scene objects. Thus the set of poses for a given surface will form a clique in an occlusion graph.

Edges between vertices in different cliques represent mutually inconsistent poses for the two surfaces. A maximum independent set of the occlusion graph will therefore contain one vertex from each clique in the graph.

A model satisfying the pose clauses selects a unique pose for each surface. The task of finding a consistent pose for every surface is therefore equivalent to selecting one vertex from each clique in the occlusion graph, which in turn is equivalent to finding the maximum independent set in the occlusion graph. □

Lemma 9.27 Let \( G = (V, E) \) be an occlusion graph. and let \( V' \subset V \) be an independent set of size \( k < |\Sigma| \), where \( |\Sigma| \) is the number of surfaces. Suppose \( s \) is a surface that has no pose in \( V' \). If \( s' \) is an arbitrary surface, then \( s \) and \( s' \) have inconsistent poses iff there is an edge \((\rho_1, \rho_2) \in E \) such that \( \rho_1 \) is a pose for \( s \), \( \rho_2 \) is a pose for \( s' \) and \( \rho_2 \in V' \).

Proof: Let \( V' \) be a maximum independent set of \( G \) of size \( k < |\Sigma| \). Suppose \( s \) is a surface that has no pose in \( V' \), and suppose \( s, s' \) have inconsistent poses. By construction of \( G \), there is an edge \((\rho_1, \rho_2) \in G \) iff \( \rho_1, \rho_2 \) are mutually inconsistent poses for two surfaces \( s, s' \). By construction of the maximum independent set, \( \rho_1 \notin V' \) iff there is a vertex \( \rho_2 \in V' \) that is adjacent to \( \rho_1 \). □

In other words, if a surface \( s \) does not have a consistent pose in \( V' \), it is because all poses that are inconsistent with the poses in the pose clause of \( s \) are already in \( V' \).

We can now use these two lemmas to provide a graph theoretical algorithm to solve the SELECT-POSE task.
Theorem 9.4 There is an $O(|R|)$ algorithm that solves the SELECT-POSE TASK for $T \cup I \cup RDA$, where $|R|$ is the number of regions in the image $I$.

Proof: The SELECT-POSE task is equivalent to finding the maximum independent set of the occlusion graph. By Lemma 9.26, every surface has a consistent pose if the maximum independent set of the occlusion graph has size $|\Sigma|$. By Lemma 9.27, the algorithm returns the set of surfaces with inconsistent poses when the size of the maximum independent set has a size less than $|\Sigma|$.

The occlusion graph can be constructed in $O(|r|)$ time. Further, there is an $O(|r|)$ algorithm to find the maximum independent set of graphs which are isomorphic to occlusion graphs ([Hutton 90]).

We can examine the maximum independent sets for the occlusion graphs in Figure 9.7 and Figure 9.8. In Figure 9.7(a) and (b), the maximum independent sets are either $\{\rho_1\}$ or $\{\rho_2\}$, so that the cardinality of the maximum independent set is less than the number of surfaces. Thus, given the assignment of surface depiction literals, there is no consistent set of poses for the two surfaces. We can see this intuitively by looking at the image – if the regions $r_1$ and $r_2$ depict different surfaces, then they must be mutually occluding, which is impossible. Thus, the algorithm SELECT-POSE returns the set $(s_1, s_2)$ of surfaces with inconsistent poses.

In Figure 9.7b, there are two maximum independent sets for the occlusion graph: $\{\rho_1, \rho_3\}$ or the set $\{\rho_2, \rho_4\}$, so that again, given the assignment of surface depiction literals, there is no consistent set of poses. The algorithm SELECT-POSE returns the sets $(s_1, s_2), (s_1, s_3)$ of surfaces with inconsistent poses.

In Figure 9.8b, there are two sets of consistent poses:

$\{\rho_5, \rho_4\}, \{\rho_5, \rho_1\}$

Thus SELECT-POSE either returns a consistent pose for every surface, or it returns a set of surfaces with inconsistent poses. Now this set of inconsistent poses arises solely from the assignment of surface depiction literals to the regions in the image. For example, in Figure 9.7(a), we have $\Delta(r_1, s_1) \land \Delta(s_2, s_2)$ but the only possible poses for $s_1$ and $s_2$ are $ontop(s_1, s_2)$ and $ontop(s_2, s_1)$. We must thus show that there is way of reassigning surface depiction literals in this case so that all surfaces can have a consistent pose.

Theorem 9.5 There is an $O(|R|^2)$ algorithm that solves the POSE-CONSISTENCY TASK for $T \cup I \cup RDA$, where $|R|$ is the number of regions in the image $I$.

Proof: By the preceding theorem, SELECT-POSE returns a consistent pose for every surface or the set of surfaces with no consistent poses. If all surfaces have consistent poses, then we are done.

If there are two surfaces that do not have consistent poses, there are two possibilities. The pose clauses for these two surfaces can be written as

$\Delta(r_i, s_i) \lor \Delta(r_j, s_j) \land (ontop(s_j, s_i) \lor s_i = s_j)$

$\Delta(r_j, s_j) \lor \Delta(r_i, s_i) \land (ontop(s_i, s_j) \lor s_i = s_j)$
SELECT-POSE(Σ, P)

Input: The pose clauses P for a set of surfaces Σ.

Output: A pose for each surface in Σ consistent with respect to T ∪ I ∪ RDA or the set of surfaces with inconsistent poses.

1. C := ∅

2. Construct a graph G = (V, E) where $V = \{\rho_i | \rho_i \text{ is a pose clause in } P\}$ and $E = \{(\rho_1, \rho_2) | \neg \rho_1 \lor \neg \rho_2 \text{ is a clause in } \Pi\}$.

3. Find the maximum independent set $V'$ of G, such that all surfaces with unique poses have a vertex in $V'$ unless two such unique poses are mutually inconsistent.

4. If $|V'| < |\Sigma|

   (a) For every surface $s \notin V'$

      i. $s' := \text{surface in } V' \text{ adjacent to } s.$

      ii. $C := C \cup \{(s, s')\}$

   (b) return C

5. else return $V'$

Figure 9.9: Algorithm for SELECT-POSE TASK
POSE-CONSISTENCY($\Delta_R, P$)

**Input**: A complete assignment $\Delta_R$ of surface depiction literals to the regions in $I$, and the set of pose clauses $P$ for all regions depicting surfaces.

**Output**: A complete assignment of surface depiction literals to the regions in $I$ consistent with respect to $T \cup I \cup RDA$ and a pose for each surface in $\Delta_R$, or INCONSISTENT.

1. $F := $ SELECT - POSE($\Delta_R, P$)

2. If $F \neq \emptyset$

   (a) For all inconsistent surface depiction literal assignments $\{\Delta(r_i, s_i), \Delta(r_j, s_j)\}$ in $F$, if $s_i \neq s_j$,

   i. $\Delta_R := \Delta_R \setminus \{\Delta(r_i, s_j)\} \cup \{\Delta(r_j, s_i)\}$

   ii. $P := $ pose clauses for regions in $\Delta_R$

   iii. POSE-CONSISTENCY($\Delta_R, P$)

(b) else Return INCONSISTENT.

3. Return $\Delta_R$.

---

**Figure 9.10**: Algorithm for the POSE-CONSISTENCY TASK

Thus, if $\Delta(r_i, s_i) \wedge \Delta(r_j, s_j) \wedge s_i \neq s_j$ is an inconsistent assignment of surface depiction literals because of inconsistent poses for $s_i, s_j$, we make the new assignment $\Delta(r_i, s_j) \wedge \Delta(r_j, s_i)$.

If on the other hand, both $\Delta(r_i, s_i) \wedge \Delta(r_j, s_j) \wedge s_i \neq s_j$ and $\Delta(r_i, s_i) \wedge \Delta(r_j, s_j) \wedge s_i = s_j$ are inconsistent, there is no consistent assignment of surface depiction literals in which both $r_i, r_j$ depict surfaces.

Reassigning surface depiction literals does not change the pose clauses for any other surfaces: this is strictly a local modification. We will therefore never retract a reassignment.

The complexity of POSE-CONSISTENCY is $O(|r|^2)$: the maximum independent set algorithm is $O(|r|)$ for a given assignment of surface depiction literals, and there is a maximum of $|r|$ reassignments of surface depiction literals. $\square$

We can see how the algorithm for POSE-CONSISTENCY works on the above examples. In Figure 9.7(a), the reassignment $\Delta(r_1, s_1) \wedge \Delta(r_2, s_1)$ is consistent: in (b), this assignment is not consistent, and in fact, there is no model for this image. In (c), there are two possible reassignments depending on which maximum independent set was given by SELECT-POSE - $\Delta(r_1, s_1) \wedge \Delta(r_2, s_1)$ or $\Delta(r_1, s_1) \wedge \Delta(r_3, s_1)$. Either of these are consistent; the first entails the pose $\Delta(r_1, s_1) \wedge ontop(s_1, s_2)$ for $s_2$, while the second entails the pose $\Delta(r_1, s_1) \wedge ontop(s_1, s_2)$ for $s_2$. In Figure 9.8(d), all surfaces have consistent poses, so that POSE-CONSISTENCY need not do any surface depiction literal reassignment. On the other hand, Figure 9.8a, no reassignment is possible, so that the given assignment of surface depiction literals
is inconsistent. In this case, one possibility is that the regions involved do not depict surfaces at all, but are merely background: because of RDA, however, such a reassignment may lead to changes in the assignment of edge depiction literals.

It should be noted that the algorithms SELECT-POSE and POSE-CONSISTENCY are essentially the only algorithms presented in this thesis. In the next chapter, we will consider the complexity of finding models of the CardWorld axioms together with various grouping assumptions. However, we will only present polynomial algorithms for tractable subclasses of these problems, and in these cases, the assignment of depiction literals is achieved trivially by straightforward application of the grouping assumptions. As we observed at the beginning of this section, once depiction literals have been assigned, we only need SELECT-POSE and POSE-CONSISTENCY to construct a model of the theory.
Chapter 10

Grouping Assumptions

In computer vision, grouping has often been cited as a way of greatly reducing the search for an interpretation of an image. However, the analysis of the computational performance of grouping has either been empirical or the formal analysis of particular algorithms. There has been no formal characterization of either the task of object recognition or the utility of grouping in object recognition. Consequently, there has been no characterization of the inherent complexity of the task. This chapter presents a formalization of the notion of grouping and explores the role it plays in the complexity of the task of object recognition. We will also show that the general problem of finding the model of an image for various grouping assumptions is NP-hard. However, we also look at sentences that define classes of images for which the task of finding a model for the grouping assumptions is tractable.

10.1 Grouping

The motivation behind grouping in image interpretation is that local geometric relations in the image can be used to identify structure in the scene. This is usually justified by some non-accidentalness principle – it is unlikely that two lines in the image satisfy some relation when the edges corresponding to these lines do not satisfy the corresponding relation in the scene. Such geometric relations include collinear \(^1\) and parallel lines, intersecting lines, and proximity.

Previous approaches (e.g. [Lowe 85], [Jacobs 88]) have considered particular arrangements of lines that are unlikely to occur at random. In this paper, the geometric relations that are used in the process of grouping are derived from the theory of surfaces and from the depiction axioms of scene relations that occur in these surfaces. An important element in any grouping strategy is that we conjecture scene relations based on image relations alone, that is, we go from image to scene. This is in contrast to the depiction axioms, which intuitively define the mapping from scene to image. Later in this section we will see how the grouping assumptions are in some sense converses and completions of the depiction axioms.

\(^1\)For readability, we will use the term collinear with the following definition:

\[
(\forall l_1, l_2) \text{collinear}(l_1, l_2) \equiv \theta^*(l_1, l_2)
\]
Previous work ([Jacobs 88]) has also been aimed at designing local algorithms that find possible groupings. The groupings generated in this way are not guaranteed to be globally consistent: it is even unclear what it means for a grouping to be consistent. We will formally define the notion of grouping in the following way:

**Definition 10.1**

\[(\forall r) \text{bgrouplng}(r) \equiv \text{region}(r) \land \neg (\exists s) \Delta(r, s)\]

\[(\forall r_1, r_2) \text{rgrouping}(r_1, r_2) \equiv \text{region}(r_1) \land \text{region}(r_2) \land (r_1 \neq r_2) \land (\exists s) (\Delta(r_1, s) \land \Delta(r_2, s))\]

**Definition 10.2** A **background grouping** is a ground literal of the form bgrouplng(r) and a region grouping is a ground literal of the form rgrouping(r_1, r_2).

A background grouping is satisfied by a model \(M\) of \(T_{cw} \cup I\) iff \(M \models \neg \exists s \Delta(r, s)\). A region grouping is satisfied by a model \(M\) of \(T_{cw} \cup I\) iff \(M \models \exists s \Delta(r_1, s) \land \Delta(r_2, s)\).

A grouping is consistent iff there is a model of \(T_{cw} \cup I\) that satisfies it.

Thus a grouping either asserts two regions depict the same surface or a region does not depict a surface. In Figure 10.1a, bgrouplng(r_1) is a background grouping, as are bgrouplng(r_1) and bgrouplng(r_2) in Figure 10.1b. In Figure 10.1c, rgrouping(r_1, r_2) is a region grouping, and in Figure 10.1d, rgrouping(r_1, r_2) and rgrouping(r_1, r_3) are region groupings.

**Definition 10.3** A grouping \(\alpha\) blocks a grouping \(\beta\) iff

\[T_{cw} \cup I \models \neg (\alpha \land \beta)\]

Thus two groupings block each other iff there is no model of \(T_{cw} \cup I\) that satisfies both \(\alpha\) and \(\beta\).

In Figure 10.1b, bgrouplng(r_1) and bgrouplng(r_2) block each other since \(l_1\) must depict an edge in some surface, forcing either \(r_1\) or \(r_2\) to depict a surface.

Note that the consistency of a grouping depends on the shape closure, that is, with different sets of allowable surfaces, different sets of groupings will be consistent. For example, consider Figure 10.1e. If the shape closure is \((\forall s) \text{surface}(s) \supset \text{rectangle}(s)\), then \(\Delta(r_1, s_1) \land \Delta(r_2, s_1)\) is not a consistent grouping, but if the shape closure is \((\forall s) \text{surface}(s) \supset \text{rectangle}(s) \lor \text{lshape}(s)\), then it is a consistent grouping.

Similarly, shape closure determines when two region groupings block each other. In Figure 10.1d, rgrouping(r_1, r_2) and rgrouping(r_1, r_3) block each other if only rectangles are allowed, but they are consistent if \(\text{Lshape}(s)\) are in the shape closure.

What is the relationship between groupings and models in the depicted kernel?

**Theorem 10.1** Let \(M\) be a non-minimal model in the depicted kernel of \(T_{cw} \cup I\). Then there exists a background grouping \(\gamma\) such that \(M \models \neg \gamma\).

**Proof:** Suppose \(M\) is a non-minimal model in the depicted kernel. Then there is a model \(M' \subset M\), and thus there is a surface \(s\) in the domain of \(M\) that is not in the domain of \(M'\). By the depicted kernel axiom, \(s\) is depicted by a region \(r\) or has an edge that is depicted. Since a region can only depict a unique surface, \(r\) cannot depict a surface in \(M'\), so that \(\Delta(r, s)\) is a background grouping that is violated in \(M'\). □
Note that the converse to this theorem is false, since minimal models can violate background groupings.

Throughout this thesis, we will be using relations among lines (such as collinear, parallel, perp) to conjecture certain scene relations. However, region groupings can be made even in the absence of accidental alignments, e.g. in Figure 10.1(f) and (g) we can group \( r_1 \) and \( r_2 \) even though there are no lines in \( r_1 \) that can depict edges in the surface depicted by \( r_2 \). This would violate the Region Depiction Assumption (RDA) introduced in the preceding chapter, which states that every region that depicts a surface has a line that depicts an edge in that surface. Thus two regions will be grouped together iff there are lines in the regions that depict edges in the same surface:

\[
T_{cw} \cup \text{RDA} \models (\forall r, r') \text{region}(r) \land \text{region}(r') \land \text{surface}(s) \supset \\
\Delta(r, s) \land \Delta(r', s) \equiv (\exists l, l', e, e') \text{in}(l, r) \land \text{in}(l', r') \land \Delta(l, e) \land \Delta(l', e') \land \text{part}(e, s) \land \text{part}(e', s))
\]

and a region will be background iff there are no such lines in the region:

\[
T_{cw} \cup \text{RDA} \models (\forall r) \text{region}(r) \supset (\exists s) \Delta(r, s) \equiv (\exists l, e) \text{in}(l, r) \land \Delta(l, e) \land \text{part}(e, s))
\]

The grouping assumptions which we present in this chapter will conjecture that two lines depict edges in the same surface; by the preceding sentences, this will entail a region grouping. Thus the grouping assumptions assign edge depiction literals to the lines in an image, and through RDA assign surface depiction literals to the regions of the image.

### 10.2 Methodology

This chapter explores the tractability of finding models for extensions of \( T_{cw} \cup I \) using different grouping assumptions which are specified as depiction sentences. Each grouping assumption specifies a different theory (the first case we consider will be the theory \( T_{cw} \cup I \), which has an empty set of grouping assumptions).

For each theory, we will consider the following:

1. Define the grouping assumption, and explore the relationship to other grouping assumptions:
2. Define classes of images related to the grouping assumption;
3. Specify a clausal theory which is equivalent to the grouping assumption;
4. Show that the complexity of finding a model of the theory for some class of images is NP-hard.
   
   Each proof will construct an image in the class and then show that the corresponding set of clauses is equivalent to the problem of PLANAR 3SAT.
5. Show that the complexity of finding a model of the theory for a more restricted class of images is polynomial.
Figure 10.1: Examples of grouping.
10.3 Ambiguous Images

Grouping assumptions construct a model of an image by assigning depiction literals to the lines and regions in the image. In this section we consider those lines for which this assignment is done without alignment-based grouping assumptions. This will lead us to formalize the notion of ambiguity in an image.

10.3.1 The Role of Ambiguous Lines

Intuitively, a line is ambiguous if we cannot assign it a depiction literal based on image properties alone. Thus, a line would be ambiguous if \( I \models in(l_1, r_1) \land in(l_1, r_2) \), but \( T_{cw} \cup I \not\models \Lambda(l_1, r_1) \) and \( T_{cw} \cup I \not\models \Lambda(l_1, r_2) \). For convex surfaces, we can identify the image properties that allow us to assign depiction literals in these cases.

**Definition 10.4**

\[
(\forall l, r) \ bounds(l, r) \equiv line(l) \land region(r) \land in(l, r) \land (\exists l', j) \ intersect(l, l', j) \\
\land (\exists a) \ \theta(l, l') \land (between(l, l', r) \equiv a < \pi) \\
\land (\forall l'') \ intersect(l, l'', j) \supset l'' = l'
\]

In other words, a line \( l \) bounds a region \( r \) iff it intersects a unique line and the region \( r \) is between the two lines iff their alignment is less than \( \pi \). Note that \( bounds \) is an image relation, and all relations in its definition are image relations.

Consider the image in Figure 10.2(a). The lines \( l_2, l_4, l_5, l_6 \) all bound the region \( r_2 \).

The utility of this definition is that lines that bound regions can be given a complete assignment of edge depiction literals. Recall that \( \Pi \) is the set of clauses equivalent to \( T_{cw} \cup I \) defined in the previous chapter.

**Lemma 10.1** If

\[ I \models in(l_1, r_1) \land in(l_1, r_2) \land bounds(l_1, r_1) \]

then

\[ \Pi \models \Lambda(l_1, r_1) \land \neg \Lambda(l_1, r_2) \]

**Proof:** By the definition of \( bounds \) there must exist a line \( l_2 \) such that

\[ I \models intersect(l_1, l_2, j_1) \land (between(l_1, l_2, r_1) \equiv \theta(a, l_1, l_2) \land a < \pi) \]

By \( RACA \), we have

\[ \Pi \models \neg \Lambda(l_1, r_2) \land \neg \Lambda(l_2, r_2) \]

so suppose

\[ \Pi \models \neg (\lambda(l_1, r_1) \land \Lambda(l_2, r_1)) \]

Then by construction of \( \Pi \), the clause equivalent to \( DTR \) (Depicted Trichotomy Assumption \(^2\)) gives

\[ \Pi \models \Delta(r_1, s_1) \land \Delta(r_2, s_2) \supset onto(s_1, s_2) \land onto(s_2, s_1) \]

\(^2\)Recall the definition of \( DTR \) from section 8.4.2.
which is inconsistent with the occlusion clauses in Π. Therefore we must have

\[ Π \models Λ(l_1, r_1) ∧ Λ(l_2, r_1) \]

and hence

\[ Π \models Λ(l_1, r_1) ∧ ¬Λ(l_1, r_2) \]

Thus, for a bound line, we can uniquely determine which surface contains the edge depicted the line.

Note that this lemma holds only for convex surfaces, since nonconvex surfaces can consistently satisfy \( ontop(s, s') ∧ ontop(s', s) \). In this case, we could consistently assign the depiction literal \( Λ(l_1, r_2) \), so that we would still have ambiguity in the image.

We can therefore define an unambiguous line with the following recursive definition:

**Definition 10.5** An unambiguous line \( l \) in a region \( r \) is one that satisfies

\[
(∀l, r) \text{ unambiguous}(l, r) \equiv \text{line}(l) ∧ \text{region}(r) ∧ \text{in}(l, r) ∧ \text{bounds}(l, r) ∧ \\
¬(∃l') \text{ in}(l, r') ∧ r \neq r' ∧ \text{bounds}(l, r') \\
∨(∃l', r'') \text{ in}(l', r'') ∧ \text{in}(l, r'') ∧ ¬(\text{parallel}(l, l') \lor \text{perp}(l, l')) ∧ \text{unambiguous}(l', r'')
\]

i.e. a line is unambiguous iff it bounds a unique region, or it is neither parallel nor perp to an unambiguous line.

The first part of the definition intuitively uses the bounds relation to determine which region depicts the surface; while in the second conjunct, we know that the lines must depict edges in different surfaces, since they are neither perp nor parallel. As an example, all lines in the image in Figure 10.2b are unambiguous.

Notice that we can decide unambiguous in \( O(|L|) \) time, where \( |L| \) is the number of lines in the image. The first disjunct in the definition provides the base case, and the second disjunct propagates the assignments of depiction literals; lines without assignments are ambiguous.

We can similarly define the notion of ambiguous lines; for such lines, there does not exist any region for they are unambiguous:

**Definition 10.6**

\[
(∀l) \text{ ambiguous}(l) \equiv \text{line}(l) ∧ (∀r) \text{ region}(r) ⊃ ¬\text{unambiguous}(l, r)
\]

We can now define classes of images based on the ambiguity in those images. \(^3\) Again, we decide membership of an image in one of these classes in time linear with respect to the lines in the image.

**Definition 10.7** The No Ambiguous Lines assumption (NAL) is the sentence

\[
(∀l) \text{ line}(l) ⊃ ¬\text{ambiguous}(l)
\]

Note that this is equivalent to \( (∀l) \text{ line}(l) ⊃ (∃r) \text{ region}(r) ∧ \text{unambiguous}(l, r) \), that is, all lines are unambiguous with respect to some region. The image in Figure 10.2b satisfies NAL.

\(^3\)In this chapter, we will restrict ourselves to planar images.
Figure 10.2: Examples for ambiguous lines.
Definition 10.8 The No Collinear Ambiguous Lines assumption (NCAL) is the sentence

\((\forall l)\ ambigious(l) \supset \neg(\exists l') \collinear(l, l')\)

In Figure 10.2c, there are three ambiguous lines \(- l_1, l_2, l_3\); since none of these lines are \collinear with any other lines, this image satisfies NCAL.

Definition 10.9 The Isolated Ambiguous Lines assumption (IAL) is the sentence

\((\forall l, l', r) \ ambigious(l) \land \in(l, r) \land \ ambigious(l') \land l \neq l' \supset \neg \in(l', r)\)

In Figure 10.2c, there are two regions \(r_1, r_2\) such that

\[ I \models \in(l_1, r_1) \land \in(l_2, r_1) \land \in(l_3, r_2) \land \ ambigious(l_1) \land \ ambigious(l_2) \]

so that this image does not satisfy IAL. In Figure 10.2d, there are only two ambiguous lines \(l_1, l_2\), and they are the only ambiguous lines in their regions: thus this image satisfies IAL. However, since \(I \models \collinear(l_1, l_2)\), this image does not satisfy NCAL.

Definition 10.10 The No Pairs of Ambiguous Lines assumption (NPAL) is the sentence

\((\forall l, l', r) \ ambigious(l) \land \ ambigious(l') \land \in(l, r) \land \in(l', r) \supset \neg(\exists l'') \collinear(l, l'') \land \neg \collinear(l', l'')\)

The following relationships are easy consequences of the definitions.

Proposition 10.1 For any image \(I\),

\[
T_{cw} \cup I \models NAL \supset NCAL \\
T_{cw} \cup I \models NCAL \supset NPAL \\
T_{cw} \cup I \models NAL \supset IAL \\
T_{cw} \cup I \models IAL \supset NPAL
\]

The relationships among the image classes is shown in Figure 10.3. The arrows represent entailment between the two sentences defining the image classes.

The important aspect of the definition of \textit{unambiguous} is that it allows us to assign depiction literals to lines and regions in the same way that \textit{bounds} allows us to uniquely assign depiction literals:

Theorem 10.2 Depiction Literal Assignment

For any planar image \(I\),

\[
T_{cw} \cup I \cup RDA \models (\forall l, r) \ unambiguous(l, r) \supset (\exists e, s) \Delta(r, s) \land \Delta(l, e) \land \part(e, s)
\]

Proof: (By induction on unambiguous lines in \(I\))

Suppose \(I \models \in(l_1, r_1) \land \bounds(l_1, r_1)\).

Base case: \(I \models \neg(\exists r) \in(l_1, r) \land r_1 \neq r \land \bounds(l_1, r)\). This is equivalent to \(I \models (\forall r) r_1 \neq r \supset \neg \in(l_1, r) \land \neg \bounds(l_1, r)\), so that there are two cases:

1. \(I \models (\forall r) \in(l_1, r_1) \land (r \neq r_1) \supset \neg \in(l_1, r)\). In this case, \(l_1\) is in a unique region. By \textit{IDA} we have

\[
T_{cw} \cup I \cup RDA \models (\exists e, s) \Delta(l_1, e) \land \part(e, s) \land \Delta(r_1, s)
\]
Figure 10.3: Special classes of images for ambiguous lines.

2. \( I \models (\forall r) in(l_1, r) \land in(l_1, r_1) \land r \neq r_1 \supset \neg \text{bounds}(l_1, r) \). In this case, \( l_1 \) is in two regions, but it only bounds one of them. Lemma 10.1 gives us

\[
T_{cw} \cup I \cup RDA \models (\forall r, s, e) r \neq r_1 \land \Delta(r, s) \land \Delta(l_1, e) \supset \neg \text{part}(e, s)
\]

\( IDA \) then entails

\[
T_{cw} \cup I \cup RDA \models (\exists e, s) \Delta(l_1, e) \land \text{part}(e, s) \land \Delta(r_1, s)
\]

Induction step: Suppose

\[
I \models in(l_2, r_2) \land \text{unambiguous}(l_2, r_2)
\]

\[
\land in(l_1, r_1) \land in(l_1, r_2) \land \neg (\text{parallel}(l_2, l_1) \lor \text{perp}(l_1, l_2))
\]

and

\[
T_{cw} \cup I \cup RDA \models (\exists e, s) \Delta(r_2, s) \land \Delta(l_2, e) \land \text{part}(e, s)
\]

i.e., we have the depiction literal assignment for an unambiguous line \( l_2 \).

\( RACA \) entails

\[
T_{cw} \cup I \cup RDA \models (\forall s, e_1, e_2) \Delta(r_2, s) \land \Delta(l_2, e_2) \land \Delta(l_1, e_1) \land \text{part}(e_2, s) \supset \neg \text{part}(e_1, s)
\]

\( IDA \) then entails

\[
T_{cw} \cup I \cup RDA \models (\exists s, e_1, e_2) \Delta(l_1, e_1) \land \text{part}(e_1, s) \land \Delta(r_1, s)
\]

\( \Box \)

For example, in Figure 10.2a, we can make the assignment \( \Delta(r_2, s_2) \), but we cannot assign \( r_1 \) or \( r_3 \) surface depiction literals, since there are ambiguous lines in these regions.

If \( I \models NAL \), then every line is assigned a depiction literal by the definition of \textit{unambiguous}; also, lines that are \textit{unambiguous} assign surface depiction literals to the region that they are in.

However, if there are ambiguous lines (i.e., \( I \nmodels NAL \)), then not every region \( r \) has a line \( l \) such that \( I \models \text{unambiguous}(l, r) \); these regions are not assigned surface depiction literals by the definition of \textit{unambiguous}. Thus we do not know whether such regions depict surfaces or not. Further, we do not know which surface contains the edge depicted by the ambiguous line. We will now show that this ambiguity leads to intractability.
10.3.2 Intractability of Ambiguous Images

For the following theorem, we will consider the problem of generating a model of $T_{cw} \cup I \cup RDA$. We will show that the problem is NP-hard using a reduction from the NP-hard problem PLANAR 3SAT: this problem was introduced in [Lichtenstein 82], and has been used in NP-hardness proofs for Blocks World in [Kirousis and Papadimitriou 88] and Mapsee in [Selman 89].

Definition 10.11 PLANAR 3SAT is a special case of satisfiability where the formulae have the following structure:

- Each clause $C_i$ has at most three literals.
- Each variable or its negation appears in at most five clauses.
- If the variables and clauses of the formulae are nodes of a graph $G$, with an edge connecting a variable and a clause whenever the variable or its negation appears in this clause, then $G$ is a planar graph.

We will use this class of formulae in all NP-hardness proofs in this chapter.

Theorem 10.3 Let $I = \{ I : I \models NCAL \}$; then for $I \in I$ it is NP-hard to find a model of $T_{cw} \cup I \cup RDA$ or determine that a model does not exist.

Proof: We will show that the problem is NP-hard using a reduction from the NP-hard problem PLANAR 3SAT. Given an instance $F$ of PLANAR 3SAT, we will construct an image $I$ such that $I \models NCAL$ and such that $F$ has a satisfying interpretation iff $T_{cw} \cup I \cup RDA$ is consistent.

Part I: CONSTRUCTION

First, given a set of PLANAR 3SAT clauses, we will construct a set $\Sigma$ of using $\Lambda(l, r)$ literals such that $\Sigma$ is satisfiable iff the original set of PLANAR 3SAT clauses is satisfiable.

In each PLANAR 3SAT clause replace each negative letter by a new letter, i.e., replace the letter $\neg p$ by $p'$. This set of clauses is reverse Horn (in fact they are positive clauses). Next add clauses that make $p'$ act like $\neg p$. Therefore add the two clauses

$$\neg p' \lor \neg p$$
$$p' \lor p$$

We will transform this propositional theory into an equivalent theory that uses $\Lambda(l, r)$ literals.

Replace each letter $p$ in a clause by the literal $\Lambda(l_i, r_j)$ for distinct $l_i$ and $r_j$.

For every letter $p$ that appears in different clauses, if $p$ has been replaced by $\Lambda(l_i, r_j)$ in one clause and has been replace by $\Lambda(l_k, r_m)$ in another clause, introduce the clauses

$$\neg \Lambda(l_i, r_j) \lor \Lambda(l_k, r_m)$$
$$\Lambda(l_i, r_j) \lor \neg \Lambda(l_k, r_m)$$

Now that we have a set of clauses $F$ using $\Lambda(l, r)$ literals, we will construct an image $I$ such that $F$ is satisfiable iff $T_{cw} \cup I \cup RDA$ is satisfiable. Note that this image will be planar, since the original instance from 3SAT is planar.

Replace each ternary clause

$$\Lambda(l_1, r_1) \lor \Lambda(l_2, r_2) \lor \Lambda(l_3, r_3)$$

by the set of lines and regions in Figure 10.4(a), where $l_1, r_1, l_2, r_2, l_3, r_3$ are parameters unique for each clause.

Finally, we introduce the following construct:
Figure 10.4: Image for representing ternary clauses in Theorem 10.3.
Definition 10.12 An ambiguous occlusion loop is a set $I$ of lines and regions $\{l_1, r_1, ..., l_n, r_n\}$ such that for all $1 < i \leq n$ we have

1. $$I \models \text{in}(l_i, r_i) \land \text{in}(l_i, r_{i-1}) \land \text{in}(l_n, r_1) \land \text{in}(l_n, r_n)$$

2. $$I \models \text{in}(l_i, r_{i-1}) \land \text{in}(l_{i-1}, r_{i-1}) \land \text{ambiguous}(l_i) \land \neg \text{collinear}(l_i, l_{i-1})$$

3. $$I \models \text{in}(l_n, r_n) \land \text{in}(l_1, r_1) \land \text{ambiguous}(l_n) \land \neg \text{collinear}(l_1, l_n)$$

An example of an ambiguous occlusion loop is the image in Figure 10.5.

For each literal $\Lambda(l_i, r_j)$ that appears in a binary and a ternary clause, $r_j$ is a region in an ambiguous occlusion loop. If $\Lambda(l_i, r_j) \equiv \Lambda(l_k, r_m)$, then $l_i, r_j$ are an even number of lines and regions away from $l_k, r_m$ in the ambiguous occlusion loop. If $\Lambda(l_i, r_j) \equiv \neg \Lambda(l_k, r_m)$, then $l_i, r_j$ are an odd number of lines and regions away from $l_k, r_m$ in the ambiguous occlusion loop.

Part II: PROOF OF EQUIVALENCE

Lemma 10.2 Consider the construct $I_1$ in Figure 10.4.

$$T_{cw} \cup I_1 \cup RDA \models \Lambda(l, r) \supset \Lambda(l_1, r_1) \lor \Lambda(l_2, r_2) \lor \Lambda(l_3, r_3)$$

Proof: The following pose clauses are satisfied by $T_{cw} \cup I_1 \cup RDA$ (these poses are illustrated in Figure 10.4(b)):

$$\Lambda(l, r) \supset \Delta(r_1, s_1) \land \text{ontop}(s_1, s) \lor \Delta(r_2, s_2) \land \text{ontop}(s_2, s) \lor \Delta(r_3, s_3) \land \text{ontop}(s_3, s)$$

Thus we get

$$T_{cw} \cup I_1 \cup RDA \models \Lambda(l, r) \supset \Lambda(l_1, r_1) \lor \Lambda(l_2, r_2) \lor \Lambda(l_3, r_3)$$

□

Lemma 10.3 For all $i < n$ a loop satisfies

$$T_{cw} \cup I \cup RDA \models \Lambda(l_i, r_i) \equiv \neg \Lambda(l_{i-1}, r_{i-1})$$

$$T_{cw} \cup I \cup RDA \models \Lambda(l_i, r_i) \equiv \text{ontop}(s_{i-1}, s_i)$$

$$T_{cw} \cup I \cup RDA \models \Lambda(l_n, r_n) \equiv \neg \Lambda(l_1, r_n)$$

$$T_{cw} \cup I \cup RDA \models \Lambda(l_i, r_i) \equiv \text{ontop}(s_1, s_n)$$

Proof: By clauses of type (2) in $\Pi$, we have for all $i < n$,

$$\Lambda(l_i, r_{i-1}) \supset \neg \Lambda(l_{i-1}, r_{i-1})$$

$$\Lambda(l_n, r_n) \supset \neg \Lambda(l_1, r_n)$$

By DTR,

$$\neg \Lambda(l_{i-1}, r_{i-1}) \supset \text{ontop}(s_{i-1}, s_i)$$
Figure 10.5: Ambiguous occlusion loop.
\[ 
\neg \Lambda(l_1, r_n) \supset \text{ontop}(s_1, s_n) \\
\text{so that we have} \\
\Lambda(l_i, r_{i-1}) \supset \text{ontop}(s_{i-1}, s_i) \\
\Lambda(l_n, r_n) \supset \text{ontop}(s_1, s_n) \\
\text{For the other direction, by } RDA \text{ we have for all } i < n, \\
\neg \Lambda(l_{i-1}, r_{i-1}) \supset \Lambda(l_i, r_{i-1}) \\
\neg \Lambda(l_1, r_n) \supset \Lambda(l_n, r_n) \\
\text{By } DTR, \text{ we have} \\
\text{ontop}(s_{i-1}, s_i) \supset \neg \Lambda(l_{i-1}, r_{i-1}) \\
\text{ontop}(s_1, s_n) \supset \neg \Lambda(l_1, r_n) \\
\text{so that we get} \\
\text{ontop}(s_{i-1}, s_i) \supset \Lambda(l_i, r_{i-1}) \\
\text{ontop}(s_1, s_n) \supset \Lambda(l_n, r_n) \\
\]

Thus occlusion loops act as "communication" constructs by passing the value of variables between the different image constructs that simulate clauses. If the literal \( \Lambda(l_i, r_i) \) is assigned true in one clause, then in any clause containing that literal it must also be also be assigned true; in any clause containing \( \neg \Lambda(l_i, r_i) \) the literal must be assigned false. The construction of the loop guarantees that these truth assignments are made consistently. Lines and regions an even distance from each other in the loop are given the same truth assignment while lines and regions an odd distance away from each other in the loop are given opposite truth assignments.

**Lemma 10.4** Loops and the construct in Figure 10.4 satisfy NCAL.

It is easy to see that the image can be built from the constructs in polynomial time and that the properties of the constructs ensure that \( T_{cw} \cup I \cup RDA \) is satisfiable iff the instance of PLANAR 3SAT is satisfiable. □

Figure 10.11 shows an example of the clause construct and associated ambiguous occlusion loops. In this image, \( r_2, r_{11}, r_{12}, r_{14} \) are in one loop, \( r_3, r_4, r_5, r_6 \) are in the second loop, and \( r_7, r_8, r_9, r_{10} \) are in the third loop. The arrows indicate the directions in which the loops continue.

### 10.3.3 Tractable Classes of Images

To find a tractable subclass, we need to remove the ambiguity of which regions depict surfaces and which are background. We will show that since NAL makes a complete assignment of edge and surface depiction literals, it is the class we are looking for.

**Lemma 10.5** If \( I \models NAL \), then

\[ 
T_{cw} \cup I \cup RDA \models (\forall r)\text{region}(r) \supset (\exists s) \Delta(r, s) \\
\]

**Proof:** This follows from URDA and Theorem 10.2. □
The following lemma tells us that if $I \models NAL$, then every line is assigned a unique positive and negative depiction literal, so that we also have a complete assignment of depiction literals for all lines in the image.

**Lemma 10.6** For any planar image $I$, if $I \models NAL$ then

$$T_{cw} \cup I \cup RDA \models (\forall l, r, e, s, s') in(l, r) \land in(l, r') \land r \neq r' \land unambiguous(l, r) \land \Delta(l, e) \land \Delta(r, s) \land \Delta(r', s') \supset \neg \text{part}(e, s').$$

**Proof:** By induction on unambiguous lines in $I$

Suppose $I \models unambiguous(l_1, r_1) \land in(l_1, r_1) \land bounds(l_1, r_1)$.

Base case: $I \models \neg (\exists r) in(l_1, r) \land r_1 \neq r \land bounds(l_1, r)$. Lemma 10.1 gives us

$$T_{cw} \cup I \cup RDA \models (\forall r, s, e) r \neq r_1 \land \Delta(r, s) \land \Delta(l_1, e) \supset \neg \text{part}(e, s).$$

Induction step: Suppose

$I \models in(l_2, r_2) \land unambiguous(l_2, r_2) \land in(l_1, r_1) \land in(l_1, r_2) \land \neg (\text{parallel}(l_2, l_1) \lor \text{perp}(l_1, l_2))$

and

$$T_{cw} \cup I \cup RDA \models (\exists e, s) \Delta(r_2, s) \land \Delta(l_2, e) \land \text{part}(e, s).$$

Then $RACA$ entails

$$T_{cw} \cup I \cup RDA \models (\forall s, e) \Delta(r_2, s) \land \Delta(l_1, e) \supset \neg \text{part}(e, s).$$

\[\square\]

**Theorem 10.4** Let $I$ be the set of images satisfying $NAL$. Then there is an $O(|r|^2)$ algorithm to find a model of $T_{cw} \cup I \cup RDA$ for some image $I \in I$ where $|r|$ is the number of regions in $I$.

**Proof:** By the preceding lemmas, $NAL$ provides a complete assignment of edge depiction literals to all lines in $I$ and surface depiction literals to all regions.

Regions are assigned surface depiction literals using the POSE-CONSISTENCY algorithm. \[\square\]

**Corollary 10.1** If $I \models NAL$ and $T_{cw} \cup I \cup RDA$ is consistent, then it has a unique model.

It is important to note that the definition of unambiguous follows directly from the axioms of our theory: the grouping done here is not an assumption. However, $NAL$ is a very restricted class of images: if we wish to tractably use the notion of grouping on a larger class of images, we will need to use additional grouping assumptions.

### 10.4 Accidental Alignments

Recall that the motivation behind grouping is the intuition that local geometric relations in the image can be used to identify structure in the scene. However, previous approaches have tended to justify grouping by some non-accidentalness principle based on probability. In this section, we define accidental alignments using the scene and depiction axioms. The grouping assumptions are then generated either from this definition or from the depiction axioms themselves.
We first define assumptions which axiomatize the intuitions of grouping, in which case there are no accidental alignments. We define sets of clauses which are equivalent to these grouping assumptions, and show that the task of finding models of these clauses is NP-hard. We then define classes of images for which this task is tractable.

10.4.1 Preliminary Definitions

Our first task will be the formalization of grouping through the notion of accidental alignment. When we presented the definition of the assumption $RDA$, we stated that

$$T_{cw} \cup RDA \models (\forall r, r', s) \text{region}(r) \land \text{region}(r') \supset \Delta(r, s) \land \Delta(r', s) \equiv (\exists l, l', e, e') \text{in}(l, r) \land \text{in}(l', r') \land \Delta(l, e) \land \Delta(l', e') \land \text{part}(e, s) \land \text{part}(e', s)$$

The point of interest here is that the two lines depict edges in the same surface. Intuitively, an accidental alignment would therefore be two lines that obey an alignment relation that exists among edges in a surface, but which do not depict edges in the same surface.

Example: Consider the images in Figure 10.6 that intuitively have accidental alignments. In (b), all models of the image satisfy

$$\text{in}(l_1, r_1) \land \text{in}(l_2, r_2) \land \text{parallel}(l_1, l_2) \land \Delta(r_1, s_1) \land \Delta(r_2, s_2) \land s_1 \neq s_2$$

$$\land \Delta(l_1, e_1) \land \Delta(l_2, e_2) \land \text{part}(e_1, s_1) \land \text{part}(e_2, s_2)$$

In (a), the image has a model $\mathcal{M}$ satisfying

$$\mathcal{M} \models \text{in}(l_1, r_1) \land \text{in}(l_2, r_2) \land \text{in}(l_2, r_2) \land \text{parallel}(l_1, l_2) \land \Delta(r_1, s_1) \land \Delta(r_2, s_2) \land s_1 \neq s_2$$

$$\land \Delta(l_1, e_1) \land \Delta(l_2, e_2) \land \Delta(l_2, e_3) \land \text{part}(e_1, s_1) \land \text{part}(e_2, s_2) \land \text{part}(e_3, s_2)$$

In this model, the two surfaces are abutting and $l_1$ depicts multiple edges; one of these edges depicts an edge in a surface different from the surface containing the edge depicted by $l_2$. □

We therefore make the following definition:

**Definition 10.13** There is an accidental alignment in an image $I$ iff

$$T_{cardworld} \cup I \models (\exists l, l', e) \text{collinear}(l, l') \land \Delta(l, e) \land \neg \Delta(l', e)$$

$$\lor (\exists l, l', e, e', s) \text{parallel}(l, l') \land \Delta(l, e) \land \Delta(l', e') \land \text{part}(e, s) \land \neg \text{part}(e', s)$$

$$\lor (\exists l, l', e, e', s) \text{perp}(l, l') \land \Delta(l, e) \land \Delta(l', e') \land \text{part}(e, s) \land \neg \text{part}(e', s)$$

i.e., there is an accidental alignment in an image $I$ iff in all models of the image there exist collinear, parallel, or perp lines which do not depict edges in the same surface.
Figure 10.6: Examples for accidental alignments.
Note that the definition of accidental alignment depends on the shape closure assumption. If we considered only equilateral triangles, then two lines at an angle of 60 degrees that did not depict edges in the same surface would not be an accidental alignment. Any scene relations that are only satisfied by unknown shapes (recall these are eliminated by SCA) would also not form accidental alignments. We will extend the notion of accidental alignment to arbitrary shapes later in the thesis.

Note that the definition of accidental alignment is a depiction property, not simply an image or scene property as in [Lowe 85] and [Jacobs 88]. Another important property is the role of semantic entailment in the definition – the sentence must be satisfied in all models of the image.

**Example:** Consider the image in Figure 10.1(e). There is no accidental alignment in this image because there is a model that satisfies the negation of the definition of accidental alignment:

\[ M_1 \models \text{in}(l_1, r_1) \land \text{in}(l_2, r_2) \land \Delta(l_1, e_1) \land \Delta(l_2, e_2) \land \Delta(r_1, s_1) \land \Delta(r_3, s_1) \land \text{part}(e_1, s_1) \land \text{part}(e_2, s_1) \]

Thus in this model, regions \( r_1 \) and \( r_3 \) depict the same surface. Note that there is a model of the image that satisfies the definition of accidental alignment.

\[ M_2 \models \text{in}(l_1, r_1) \land \text{in}(l_2, r_3) \land \Delta(l_1, e_1) \land \Delta(l_2, e_2) \land \Delta(r_1, s_1) \land \Delta(r_3, s_2) \land s_1 \neq s_2 \land \text{part}(e_1, s_1) \land \text{part}(e_2, s_2) \]

Any grouping assumptions must eliminate this model of the image, so that only \( M_1 \) is consistent.

To eliminate models that satisfy accidental alignments, the first grouping assumption we will consider is the negation of the definition of accidental alignment.

**Definition 10.14** The Non Accidental Alignments (NAA) assumption is the conjunction of the following sentences:

\[ (\forall l, l', e) \text{collinear}(l, l') \land \Delta(l, e) \supset \Delta(l', e) \]

\[ (\forall l, l', e, e', s) \text{parallel}(l, l') \land \Delta(l, e) \land \Delta(l', e') \land \text{part}(e, s) \supset \text{part}(e', s) \]

\[ (\forall l, l', e, e', s) \text{perp}(l, l') \land \Delta(l, e) \land \Delta(l', e') \land \text{part}(e, s) \supset \text{part}(e', s) \]

There are other forms of nonaccidental alignment assumptions that we can consider. The other grouping assumption that we will consider in this thesis is the predicate completion ([Clark 78]) of the depiction axioms for alignment predicates. In this approach, scene relations are conjectured as explanations of the image relations, as in abduction ([Poole 87]).

**Definition 10.15** Abductive NAA (ANAA) is the conjunction of the following sentences:

\[ (\forall l, l') \text{collinear}(l, l') \supset (\exists e) \Delta(l, e) \land \Delta(l', e) \]

\[ (\forall l, l') \text{parallel}(l, l') \supset (\exists e, e', s) \Delta(l, e) \land \Delta(l', e') \land \text{part}(e, s) \land \text{part}(e', s) \]

\[ (\forall l, l') \text{perp}(l, l') \supset (\exists e, e', s) \Delta(l, e) \land \Delta(l', e') \land \text{part}(e, s) \land \text{part}(e', s) \]
It is easy to see that we have the following

**Theorem 10.5** For any planar image $I$,

$$T_{cardworld} \cup I \models \text{NAA} \supset \text{ANAA}$$

Note, however, that the converse does not hold.

**Example:** Consider Figure 10.6d; this image has a model $\mathcal{M}$ that satisfies ANAA but falsifies NAA:

$$\mathcal{M} \models \Delta(l_1, e_1) \land \Delta(l_2, e_2) \land \Delta(l_1, e_2) \land \text{parallel}(e_1, e_2)$$

$$\land \text{part}(e_1, s_1) \land \text{part}(e_2, s_1) \land \text{part}(e_3, s_2) \land \Delta(r_1, s_1) \land \Delta(r_2, s_2)$$

In this model, $l_1$ and $l_2$ are in accidental alignment because $l_2$ depicts an edge that is not part of surface $s_1$; however, both lines do depict edges in the same surface. □

We can therefore inquire as to what additional assumptions N.A.A makes that AN.A.A is not committed to. To do this we first consider the assignment of depiction literals.

The grouping assumptions assign edge depiction literals to the lines in an image $I$ such that for every line $l$ such that $l \models \text{in}(l, r_1) \land \text{in}(l, r_2)$ there are two edge depiction literals: $\Lambda(l, r_1)$ or $\neg\Lambda(l, r_1)$ and $\Lambda(l, r_2)$ or $\neg\Lambda(l, r_2)$. This assignment also provides an assignment of surface depiction literals to the regions of $I$ in the following way: if $\Lambda(l, r_1)$ then $\Delta(r_1, s_1)$; if for every line $l$ such that $l \models \text{in}(l, r)$ we have $\neg\Lambda(l, r)$, we have $\neg\Delta(r, z)$ by RDA. However, not all lines and regions can be assigned depiction literals using the grouping assumptions.

**Definition 10.16** Given an image $I$, an orphan $l$ is a line in $I$ that satisfies

$$(\forall l) \text{orphan}(l) \equiv \neg(\exists l') \text{line}(l') \land \text{parallel}(l, l') \lor \text{perp}(l, l')$$

i.e., an orphan is a line which is neither parallel nor perp to any other line in the image.

Notice that the concept of orphan depends on the shape closure assumption and the shape dependent alignment relations satisfied by the surfaces in the closure, such as parallel and perp. Thus collinear lines may be orphans.

The notion of orphan is important precisely because it allows us to assign depiction literals, analogously to the role played by ambiguous lines in the previous section. If the edges depicted by two non-orphans $l, l'$ are in the same surface, then we can assign surface depiction literals to the regions containing the lines using the depiction axioms in $T_{\Delta}$. We cannot do this with collinear lines: even if we know that two lines depict the same edge, we may not know which region depicts the surface containing the edges.

**Example:** In Figure 10.6(e), we have

$$T_{cw} \cup I \cup \text{RDA} \cup \text{NAA} \models (\forall e, e', s) \Delta(l_1, e) \land \Delta(l_2, e') \land \text{part}(e, s) \supset \text{part}(e', s)$$
However, we do not know which regions depict the surface containing the edges: there is one model where \( r_1, r_2 \) depict the surface and another model where \( r_2, r_4 \) depict the surface. On the other hand, in Figure 10.6(f), we have

\[
T_{cw} \cup I \cup RDA \cup NAA \models \Delta(l_1, e_1) \land \Delta(l_2, e_2) \land \Delta(r_3, s_1) \land \Delta(r_2, s_1) \land \Delta(r_4, s_3) \land \Delta(r_4, s_4) \\
\land \text{part}(e_1, s_1) \land \text{part}(e_2, s_1) \land \neg \text{part}(e_1, s_3) \land \neg \text{part}(e_2, s_4)
\]

Thus \( NAA \) gives a complete assignment of edge and surface depiction literals for this image. □

Using the notion of orphan, we can also see the difference between these two grouping assumptions. \( NAA \) forces non-orphans and collinear orphans to depict unique edges, since a line depicting multiple edges satisfies the definition of accidental alignment. On the other hand, \( NAA \) allows some non-orphans to depict multiple edges, as we saw in Figure 10.6d. We therefore introduce the following assumption:

**Definition 10.17** The Uniquely Depicting Multiple Lines (UDML) assumption is the sentence

\[
(\forall l, e, s, l', e', e'') \Delta(l, e) \land \Delta(l', e') \land \text{part}(e, s) \land \text{part}(e', s) \land l \neq l' \land \Delta(l, e'') \supset e = e''
\]

Thus if a surface has multiple lines depicting its edges, then those lines must be uniquely depicting. Note that this assumption applies only to multiple lines depicting edges in a surface, including non-orphans and collinear orphans. \( UDML \) forces these lines to depict unique edges. On the other hand, orphans for which there is no collinear line can depict multiple edges.

**Example:** Consider the image in Figure 10.6(c), which has a model satisfying

\[
M \models \text{orphan}(l_1) \land \Delta(l_1, e_1) \land \Delta(l_2, e_2) \land \Delta(r_1, s_1) \land \Delta(r_2, s_2) \land \text{part}(e_1, s_1) \land \text{part}(e_2, s_2)
\]

Notice that this image has no accidental alignments. □

It is easy to see that this new assumption \( UDML \) is weaker than \( NAA \):

**Theorem 10.6**

\[
T_{cw} \cup RDA \models NAA \supset UDML
\]

**Proof:** Suppose \( UDML \) is falsified

\[
T_{cw} \cup RDA \models \exists l_1, l_2, e_1, e_2, e_3, s_1, s_2, s_3, s_4 \Delta(l_1, e_1) \land \Delta(l_2, e_2) \land \Delta(l_1, e_3) \land \neg \text{part}(e_3, s_1)
\]

By \( RACA \) we must have

\[
T_{cw} \cup RDA \models \forall l_1, l_2, e_1, e_2, e_3, s_1, s_2, s_3, s_4 \Delta(l_1, e_1) \land \Delta(l_2, e_2) \land \Delta(l_1, e_3) \land \neg \text{part}(e_3, s_1)
\]

parallel(l_1, l_2) \lor \text{perp}(l_1, l_2) \lor \text{collinear}(l_1, l_2)

so that \( NAA \) entails

\[
T_{cw} \cup RDA \models \forall l_1, l_2, e_1, e_2, e_3, s_1, s_2, s_3, s_4 \Delta(l_1, e_1) \land \Delta(l_2, e_2) \land \text{part}(e_1, s_1) \land \text{part}(e_2, s_2) \land \Delta(l_1, e_3) \land \Delta(l_1, e_3)
\]

\[
\supset \text{part}(e_3, s_1)
\]

which contradicts our hypothesis. □
In fact, $NAA$ is equivalent to the conjunction of $ANAA$ and $UDML$:

**Theorem 10.7**

$$T_{cw} \cup RDA \models NAA \equiv ANAA \land UDM L$$

**Proof:** One direction follows from the previous two theorems.

For the other direction, we will show

$$T_{cw} \cup RDA \models ANAA \land \neg NAA \lor \neg UDM L$$

Suppose $NAA$ is falsified, so that we have an accidental alignment:

$$T_{cw} \cup RDA \models (\exists l_1, l_2, e_1, e_2, s) (\text{parallel}(l_1, l_2) \lor \text{perp}(l_1, l_2) \lor \text{collinear}(l_1, l_2))$$

$$\land \Delta(l_1, e_1) \land \Delta(l_2, e_2) \land \text{part}(e_1, s) \land \neg \text{part}(e_2, s)$$

However, since $ANAA$ is satisfied, there must exist an edge such that

$$T_{cw} \cup RDA \models (\exists l_1, l_2, e_3, s) \Delta(l_2, e_3) \land \Delta(l_2, e_2) \land \text{part}(e_3, s)$$

which falsifies $UDML$. $\square$

We now have a complete characterization of the two grouping assumptions used in this chapter. This distinction between the two grouping assumptions will play a crucial role in the complexity proofs later in this chapter.

**10.4.2 Classes of Images for Accidental Alignments**

We can now proceed along the same path we pursued in the previous section with ambiguous lines, and define different classes of images. In the previous section, these classes were defined with respect to the existence of ambiguous lines in the image; in this section, the classes of images will be based on the existence of orphans.

**Definition 10.18** An ambiguous orphan $l$ is a line that satisfies

$$(\forall l) \text{ambig-orphan}(l) \equiv \text{orphan}(l) \land \text{ambiguous}(l) \land$$

$$[\neg (\exists r, l', l'') \text{in}(l, r) \land \text{in}(l', r) \land (\text{parallel}(l', l'') \lor \text{perp}(l', l''))]$$

i.e., the line is ambiguous, it is an orphan, and there are no other lines in the region containing $l$ that are parallel or perp with each other.

We can define classes of images using this definition.

**Definition 10.19** The No Ambiguous Orphans assumption (NAO) is the sentence

$$(\forall l) \text{line}(l) \supset \neg \text{ambig-orphan}(l)$$

i.e., there do not exist any ambiguous orphans in the image.

The images in Figure 10.7(a) and Figure 10.7(d) satisfy NAO. Note that we have $I \models \text{ambiguous}(l_2)$ in Figure 10.7(a) and $I \models \text{ambiguous}(l_1) \land \text{ambiguous}(l_2) \land \text{ambiguous}(l_3)$. 
Definition 10.20 The No Pairs of Ambiguous Orphans assumption (NPAO) is the sentence
\[(\forall l, l', l'') \text{ambig.\ orphan}(l) \land \text{ambig.\ orphan}(l') \land \text{in}(l, r) \land \text{in}(l', r) \supset \neg (\text{collinear}(l, l'') \lor \text{collinear}(l', l''))\]
i.e., if there are any ambiguous orphans in the image, then they are not collinear with any other lines in the image.

Definition 10.21 The No Collinear Ambiguous Orphans assumption (NCAO) is the sentence
\[(\forall l, l') \text{ambig.\ orphan}(l) \supset \neg \text{collinear}(l, l')\]
i.e., if \(l\) is an ambiguous orphan, then it is not collinear with any other line in the image.

In Figure 10.7(b), there are two ambiguous orphans \(- l_1, l_2\); since none of these lines are collinear with any other lines, this image satisfies NCAO.

Definition 10.22 The Isolated Ambiguous Orphans assumption (IAO) is the sentence
\[(\forall l, l', r) \text{ambig.\ orphan}(l) \land \text{in}(l, r) \land \text{ambig.\ orphan}(l') \land l \neq l' \supset \neg \text{in}(l', r)\]
i.e., there is at most one ambiguous line in each region.

In Figure 10.7(b), there are two regions \(r_1, r_2\) such that
\[I \models \text{in}(l_1, r_1) \land \text{in}(l_2, r_1) \land \text{in}(l_2, r_2) \land \text{in}(l_3, r_2) \land \text{ambiguous}(l_1) \land \text{ambiguous}(l_2) \land \text{orphan}(l_1) \land \text{orphan}(l_2)\]
so that this image does not satisfy IAO. In Figure 10.7(c), there are only two ambiguous orphans \(l_1, l_2\), and they are the only ambiguous orphans in their regions; thus this image satisfies IAO. However, since \(I \models \text{collinear}(l_1, l_2)\), this image does not satisfy NCAO.

The following relationships are easy consequences of the definitions:

**Proposition 10.2** For any image \(I\),

\[T_{cw} \cup I \models NAO \supset IAO\]
\[T_{cw} \cup I \models IAO \supset NCAO\]
\[T_{cw} \cup I \models NAO \supset IAO\]
\[T_{cw} \cup I \models IAO \supset NPAO\]

The relationship between these classes is shown in Figure 10.8.

It is important to note that we will be constructing models of \(T_{cw} \cup I \cup RDA \cup NAA\) and \(T_{cw} \cup I \cup RDA \cup ANAA\). Thus, if a model has accidental alignments, then it will be inconsistent with these grouping assumptions. Since the images in Figure 10.1(d) and Figure 10.6(a),(b) have accidental alignments, they will be inconsistent. An alternative approach would be to construct models of images that have accidental alignments, but in which we prefer those models that minimize the accidentals. This approach will not be taken in this thesis.

We will now address the problem of constructing models satisfying the grouping assumptions.
Figure 10.7: Examples for image classes related to accidental alignments.
Figure 10.8: Special classes of images for accidental alignments.
10.4.3 Computational Framework for N.AA

In the previous chapter, we converted the first-order axioms of CardWorld into a logically equivalent set of propositional clauses. In this section, we will convert N.AA into a logically equivalent set of propositional clauses given an image $I$. This set of clauses will be used in the complexity proofs and in algorithms to construct models satisfying N.AA.

Definition 10.23 $\Pi_{N.AA}$ is the following set of clauses:

1. For all lines $l_i, l_j$ such that

$$I \models \text{in}(l_i, r_k) \land \text{in}(l_j, r_m) \land \text{in}(l_i, r_n) \land \text{in}(l_j, r_p) \land \text{between}(l_i, l_j, r_m) \land \text{between}(l_i, l_j, r_k)$$

and $I \models \text{parallel}(l_i, l_j)$ or $I \models \text{perp}(l_i, l_j)$, introduce the literals

$$\Lambda(l_i, r_k), \Lambda(l_i, r_m), \Lambda(l_j, r_k), \Lambda(l_j, r_m), \Delta(r_i, s_i), \Delta(r_j, s_i) \in \Pi_{N.AA}$$

2. If

$$I \models \text{in}(l_i, r_i) \land \text{in}(l_2, r_j) \land \text{in}(l_1, r_k) \land \text{in}(l_2, r_m) \land \text{collinear}(l_1, l_2) \land \text{orphan}(l_1) \land \text{orphan}(l_2)$$

$$\land \text{between}(l_1, l_2, r_k) \land \text{between}(l_2, l_1, r_j)$$

introduce the clauses

$$\Lambda(l_1, r_k) \lor \Lambda(l_2, r_j),$$

$$\Lambda(l_1, r_k) \lor \Delta(r_i, s_i), \Lambda(l_1, r_k) \lor \Delta(r_j, s_i),$$

$$\Lambda(l_2, r_j) \lor \Delta(r_k, s_k), \Lambda(l_2, r_j) \lor \Delta(r_m, s_k) \in \Pi_{N.AA}$$

3. If

$$I \models \text{in}(l_i, r_i) \land \text{in}(l_2, r_j) \land \text{in}(l_1, r_k) \land \text{in}(l_2, r_m) \land \text{collinear}(l_1, l_2) \land \text{orphan}(l_1) \land \text{orphan}(l_2)$$

$$\land \neg \text{between}(l_1, l_2, r_i) \land \neg \text{between}(l_2, l_1, r_m)$$

introduce the clauses

$$\Lambda(l_1, r_i) \lor \Lambda(l_2, r_m),$$

$$\Lambda(l_1, r_i) \lor \Delta(r_k, s_k), \Lambda(l_1, r_i) \lor \Delta(r_m, s_k).$$

$$\Lambda(l_2, r_m) \lor \Delta(r_i, s_i), \Lambda(l_2, r_m) \lor \Delta(r_j, s_i) \in \Pi_{N.AA}$$

4. If

$$I \models \text{in}(l_i, r_i) \land \text{in}(l_1, r_k) \land (\text{parallel}(l_1, l_2) \lor \text{perp}(l_1, l_2) \lor \text{collinear}(l_1, l_2))$$

introduce the clause

$$\neg\Lambda(l_1, r_i) \lor -\Lambda(l_1, r_k) \in \Pi_{N.AA}$$

Theorem 10.8 Let $\Phi$ be a sentence in $L(\Pi)$. Given a planar image $I$, $\Pi \cup \Pi_{N.AA} \models \Phi$ iff $T_{cw} \cup I \cup RDA \cup N.AA \models \Phi$.

Proof: In the previous chapter, we proved that the clauses in $\Pi$ are equivalent to $T_{cw} \cup RDA \cup I$ for some image $I$. We simply need to show that N.AA is equivalent to the clauses in $\Pi_{N.AA}$ for some image $I$.

1. With N.AA and CFGA, depiction axiom 9.5 of $T_{cw}$ is equivalent to

$$(\forall l, l', r, a) \text{in}(l, r) \land \text{between}(l, l', r) \land \theta(a, l, l') \land a < \pi \supset \Lambda(l, r) \lor \Lambda(l', r)$$

This gives us the literals

$$\Lambda(l_i, r_k), \Lambda(l_i, r_m), \Lambda(l_j, r_k), \Lambda(l_j, r_m), \Delta(r_i, s_i), \Delta(r_j, s_i)$$
2. With A.N.A.A., depiction axiom 9.6 of \( T_{cw} \) is equivalent to
\[
(\forall l, l', r, r') (\text{between}(l, l', r) \equiv \text{between}(l', l, r')) \supset \Lambda(l, r) \lor \Lambda(l', r')
\]
It is easy to see how this leads to the clauses in sets 2 and 3 above.

3. Suppose
\[
I \models \text{in}(l_1, r_1) \land \text{in}(l_1, r_k) \land (\text{parallel}(l_1, l_2) \lor \text{perp}(l_1, l_2) \lor \text{collinear}(l_1, l_2))
\]
By A.N.A.A. we must have
\[
T_{cw} \cup \text{RDA} \cup I \models (\exists e_1, e_2, s) \Delta(l_1, e_1) \land \Delta(l_2, e_2) \land \text{part}(e_1, s) \land \text{part}(e_2, s)
\]
By I.D.A. we must have
\[
T_{cw} \cup \text{RDA} \cup I \models (\exists s) \Delta(r, s) \land \Delta(r_k, s)
\]
so that in this case U.D.M.L is equivalent to
\[
(\forall l, l') \text{in}(l, r) \land \text{in}(l, r') \supset (\Lambda(l, r) \lor \neg \text{in}(l, r'))
\]
\( \square \)

There is one class of clauses that do not appear in \( \Pi \cup \Pi_{N.A.A.} \), namely the non-binary Horn clauses generated by sets of lines that have no consistent pose:
\[
\neg \Lambda(l_1, r_1) \lor \neg \Lambda(l_2, r_1) \lor \neg \Lambda(l_3, r_1)
\]
The following theorem shows that these clauses are inconsistent with \( N.A.A. \): in a sense, such lines are in accidental alignment since they cannot depict edges in the same surface.

**Theorem 10.9** Suppose \( T_{cw} \cup I \cup \text{RDA} \cup N.A.A. \not\models \neg (\Lambda(l_i, r_m) \land \Lambda(l_j, r_m)) \) and \( I \models \text{parallel}(l_i, l_j) \lor \text{perp}(l_i, l_j) \lor \text{collinear}(l_i, l_j) \). Then \( T_{cw} \cup I \cup \text{RDA} \cup N.A.A. \models \Lambda(l_i, r_m) \land \Lambda(l_j, r_m) \).

**Proof:** Suppose there is a model \( \mathcal{M} \) of \( T_{cw} \cup I \cup \text{RDA} \cup N.A.A. \) such that
\[
\mathcal{M} \models \neg (\Lambda(l_i, r_m) \land \Lambda(l_j, r_m))
\]
Then
\[
\mathcal{M} \models \neg (\exists e, e') \Delta(l_i, e) \land \Delta(l_j, e') \land (\text{parallel}(e, e') \lor \text{perp}(e, e') \lor e = e')
\]
But then \( l_i, l_j \) are in accidental alignment and \( \mathcal{M} \) falsifies \( N.A.A. \). \( \square \)

In other words, every grouping consistent with \( T_{cw} \cup I \cup \text{RDA} \cup N.A.A. \) is entailed by \( T_{cw} \cup I \cup \text{RDA} \cup N.A.A. \). As a result of this theorem, there are no minimal clauses of the form
\[
\neg \Lambda(l_i, r_m) \lor \neg \Lambda(l_j, r_m) \lor \neg \Lambda(l_k, r_m)
\]
in \( \Pi \cup \Pi_{N.A.A.} \). Suppose there were such a minimal clause; then we have
\[
T_{cw} \cup I \cup \text{RDA} \cup N.A.A. \not\models \neg (\Lambda(l_i, r_m) \land \Lambda(l_j, r_m))
\]
\[
T_{cw} \cup I \cup \text{RDA} \cup N.A.A. \not\models \neg (\Lambda(l_i, r_m) \land \Lambda(l_k, r_m))
\]
so that by the preceding theorem we have

\[ T_{cw} \cup I \cup RDA \cup NAA \not\models -\Lambda(l_j, r_m) \land \Lambda(l_k, r_m) \]

which is inconsistent with the clause. Thus the image in Figure 9.6 has an accidental alignment, since it satisfies

\[-\Lambda(l_{29}, r_9) \lor -\Lambda(l_{31}, r_9) \lor -\Lambda(l_{38}, r_9)\]

We will now consider the complexity of finding models of \( T_{cw} \cup I \cup RDA \cup NAA \).

**Theorem 10.10**  Let \( I = \{ I : I \models IAO \} \); then for \( I \in I \) it is NP-hard to find a model of \( T_{cw} \cup I \cup RDA \cup NAA \) or determine that a model does not exist.

**Proof:** We will show that the problem is NP-hard using a reduction from PLANAR 3SAT. Given an instance \( F \) of PLANAR 3SAT, we will construct an image \( I \) such that \( I \models IAO \) and such that \( F \) has a satisfying interpretation iff \( T_{cw} \cup I \cup RDA \cup NAA \) is consistent. Note that this image will be planar, since the original instance from 3SAT is planar.

**Part I: THE CONSTRUCTION**

We transform a PLANAR 3SAT set of clauses into an equivalent theory that uses \( \Lambda(l, r) \) literals with the same construction we used in Theorem 10.3.

**Definition 10.24** A double loop is a set \( I \) of two sequences of regions and lines

\[ \{l_0, r_1, r_2, \ldots, l_n, r_{2n}, r_{2n-1}\}, \{l'_1, r'_3, \ldots, l'_n, r'_n, r'_{n-1}\} \]

such that for all \( i \leq n \) we have

1. \( I \models orphan(l_{2i}) \land orphan(l'_{2i}) \)

2. \( I \models in(l_{2i}, r_{2i-1}) \land in(l_{2i}, r_{2i-2}) \land in(l'_{2i}, r'_{2i-1}) \land in(l'_{2i}, r'_{2i-2}) \)

3. If \( i \) is odd,

\( I \models collinear(l_{2i}, l_{2i-2}) \land collinear(l'_{2i}, l'_{2i-2}) \)

4. If \( i \) is even,

\( I \models -collinear(l_{2i}, l_{2i-2}) \land -collinear(l'_{2i}, l'_{2i-2}) \)

5. \( I \models -collinear(l_{2n-2}, l_0) \land -collinear(l'_{2n-2}, l_0) \)

6. \( I \models -collinear(l_2, l'_4) \land between(l_2, l'_4, r_4) \)
Figure 10.9: Image for representing ternary clauses in Theorem 10.10.
7. If \( I \models \text{collinear}(l_{2i}, l_{2i-2}) \), then
\[ I \models \text{between}(l_{2i}, l_{2i-2}, r_{2i-2}) \land \neg \text{between}(l_{2i-2}, l_{2i}, r_{2i-3}) \]

8. If \( I \models \neg \text{collinear}(l_{2i}, l_{2i-2}) \), then
\[ I \models \neg \text{between}(l_{2i}, l_{2i-2}, r_{2i-1}) \land \text{between}(l_{2i-2}, l_{2i}, r_{2i-2}) \]

9. If \( I \models \text{collinear}(l'_{2i}, l'_{2i-2}) \), then
\[ I \models \text{between}(l'_{2i}, l'_{2i-2}, r'_{2i-2}) \land \neg \text{between}(l'_{2i-2}, l'_{2i}, r'_{2i-3}) \]

10. If \( I \models \neg \text{collinear}(l'_{2i}, l'_{2i-2}) \), then
\[ I \models \neg \text{between}(l'_{2i}, l'_{2i-2}, r'_{2i-1}) \land \text{between}(l'_{2i-2}, l'_{2i}, r'_{2i-2}) \]

11. 
\[ T_{cv} \cup I \cup RDA \models \Delta(r_{2i-1}, s_{2i-1}) \supset \Delta(r_{2i-3}, s_{2i-3}) \land \text{ontop}(s_{2i-3}, s_{2i-1}) \]
\[ T_{cv} \cup I \cup RDA \models \Delta(r'_{2i-1}, s'_{2i-1}) \supset \Delta(r'_{2i-3}, s'_{2i-3}) \land \text{ontop}(s'_{2i-3}, s'_{2i-1}) \]

12. 
\[ T_{cv} \cup I \cup RDA \models \Delta(r_{2n-2}, s_{2n-2}) \supset \Delta(r_1, s_1) \land \text{ontop}(s_{2n-2}, s_1) \]
\[ T_{cv} \cup I \cup RDA \models \Delta(r'_{2n-2}, s'_{2n-2}) \supset \Delta(r_2, s_2) \land \text{ontop}(s', s_{2n-2}, s_2) \]
\[ T_{cv} \cup I \cup RDA \models \Delta(r_4, s_5) \supset \Delta(r'_5, s'_5) \land \text{ontop}(s'_5, s_5) \]

An example of a double loop is in Figure 10.10.

Replace each ternary clause
\[ \neg \Lambda(l_1, r_1) \lor \Lambda(l_2, r_2) \lor \Lambda(l_3, r_3) \]
by the set of lines and regions in Figure 10.9a, where \( l_1, r_1, l_2, r_2, l_3, r_3 \) are parameters unique for each clause.

For each literal \( \Lambda(l_i, r_j) \) that appears in binary and ternary clause, \( r_j \) is a region in a double loop. If \( \Lambda(l_1, r_j) \equiv \Lambda(l_k, r_m) \), then \( l_1, r_j \) are in the same sequence of the double loop. If \( \Lambda(l_i, r_j) \equiv \neg \Lambda(l_k, r_m) \), then \( l_i, r_j \) are in different sequences of the double loop.

**Part II: PROOF OF EQUIVALENCE**

**Lemma 10.7** The construct \( I_1 \) in Figure 10.4a satisfies:
\[ T_{cv} \cup I_1 \cup NAA \cup RDA \models \neg \Lambda(l_1, r_1) \lor \Lambda(l_2, r_2) \lor \Lambda(l_3, r_3) \]

**Proof:** Suppose we have a model \( \mathcal{M} \) of \( T_{cv} \cup I_1 \cup NAA \cup RDA \) such that
\[ \mathcal{M} \models \Lambda(l_1, r_1) \land \Delta(r_1, s_1) \]

Then there are only two consistent poses for \( s_1 \): 
\[ \Delta(r_2, s_2) \land \text{ontop}(s_2, s_1) \]

or
\[ \Delta(r_3, s_3) \land \text{ontop}(s_3, s_1) \]

\footnote{Figure 10.9b illustrates the two poses for \( s_1 \).}
Figure 10.10: Image for representing the double loop construct.
By RDA, one of the lines in these regions must depict an edge in those surfaces: by construction of $\Pi \cup \Pi_{NAA}$, since $l_2, l_3$ are the only orphans in their respective regions, they must be the lines that depict the edges, so that we have

$$\Lambda(l_2, r_2) \lor \Lambda(l_3, r_3)$$

Therefore,

$$T_{cw} \cup I_1 \cup NAA \cup RDA \models -\Lambda(l_1, r_1) \lor \Lambda(l_2, r_2) \lor \Lambda(l_3, r_3)$$

\[ \square \]

**Lemma 10.8** For all $i \leq n$, a double loop $I$

\[ \{l_0, r_1, r_2, \ldots, l_n, r_{2n}, r_{2n-1}\}, \{l'_4, r'_5, r'_6, \ldots, l'_n, r'_{2n}, r'_{2n-1}\} \]

satisfies

$$T_{cw} \cup I \cup NAA \cup RDA \models \Lambda(l_{2i}, r_{2i-1}) \equiv \Lambda(l_0, r_1)$$

$$T_{cw} \cup I \cup NAA \cup RDA \models \Lambda(l'_{2i}, r'_{2i-1}) \equiv \Lambda(l_0, r_2)$$

$$T_{cw} \cup I \cup NAA \cup RDA \models \Lambda(l_{2i}, r_{2i-1}) \equiv -\Lambda(l'_{2i}, r'_{2i-1})$$

**Proof:** 1. By construction of $\Pi_{NAA}$, if $I \models \text{collinear}(l_{2i}, l_{2i-2})$, we have

$$T_{cw} \cup I \cup NAA \cup RDA \models \Lambda(l_{2i}, r_{2i-1}) \equiv \Lambda(l_{2i-2}, r_{2i-3})$$

By RDA and condition (7) in the construction, if $I \models \text{collinear}(l_{2i}, l_{2i-2})$ we have,

$$T_{cw} \cup I \cup NAA \cup RDA \models \Lambda(l_{2i}, r_{2i-1}) \equiv \Lambda(l_{2i-2}, r_{2i-3})$$

and

$$T_{cw} \cup I \cup NAA \cup RDA \models \Lambda(l_{2n-2}, r_{2n-2}) \equiv \Lambda(l_0, r_1)$$

Combining these three sentences gives us

$$T_{cw} \cup I \cup NAA \cup RDA \models \Lambda(l_{2i}, r_{2i-1}) \equiv \Lambda(l_0, r_1)$$

2. Similarly, by construction of $\Pi_{NAA}$, if $I \models \text{collinear}(l_{2i}, l_{2i-2})$, we have

$$T_{cw} \cup I \cup NAA \cup RDA \models \Lambda(l'_{2i}, r'_{2i-1}) \equiv \Lambda(l'_{2i-2}, r'_{2i-3})$$

By RDA and condition (7) in the construction, if $I \models \text{collinear}(l_{2i}, l_{2i-2})$ we have,

$$T_{cw} \cup I \cup NAA \cup RDA \models \Lambda(l'_{2i}, r'_{2i-1}) \equiv \Lambda(l'_{2i-2}, r'_{2i-3})$$

and

$$T_{cw} \cup I \cup NAA \cup RDA \models \Lambda(l'_{2n-2}, r'_{2n-2}) \equiv \Lambda(l_0, r_1)$$

which combine to give

$$T_{cw} \cup I \cup NAA \cup RDA \models \Lambda(l'_{2i}, r'_{2i-1}) \equiv \Lambda(l_0, r_2)$$

3. By construction of $\Pi_{NAA}$ we have

$$T_{cw} \cup I \cup NAA \cup RDA \models \Lambda(l_0, r_1) \equiv -\Lambda(l_0, r_2)$$

so that combining the previous two results we get

$$T_{cw} \cup I \cup NAA \cup RDA \models \Lambda(l_{2i}, r_{2i-1}) \equiv -\Lambda(l'_{2i}, r'_{2i-1})$$

\[ \square \]
Thus double loops act as "communication" constructs by passing the value of variables between the different image constructs that simulate clauses. If the literal \( \Lambda(l_i, r_j) \) is assigned true in one clause, then in any clause containing that literal it must also be also be assigned true: in any clause containing \( \neg \Lambda(l_i, r_j) \) the literal must be assigned false. The construction of the double loop guarantees that these truth assignments are made consistently. Depiction literals for all lines \( l_j \) in the same loop are given the same truth assignment, while depiction literals for all lines \( l_j \) in different loops are given opposite truth assignments.

**Lemma 10.9** Double loops and the construct \( I_i \) in Figure 10.4 satisfy \( IAO \).

It is easy to see that the image can be built from the constructs in polynomial time and that the properties of the constructs ensure that \( T_{cw} \cup I \cup N.A.A. \cup R.D.A. \) is satisfiable iff the instance of 3SAT is satisfiable. □

**Theorem 10.11** Let \( I = \{ I : I \models NCAO \} \); then for \( I \in I \) it is NP-hard to find a model of \( T_{cw} \cup I \cup R.D.A. \cup N.A.A. \) or determine that a model does not exist.

**Proof:** We will show that the problem is NP-hard using a reduction from PLANAR 3SAT. Given an instance \( F \) of PLANAR 3SAT, we will construct an image \( I \) such that \( F \) has a satisfying interpretation iff \( T_{cw} \cup I \cup R.D.A. \cup N.A.A. \) is consistent. Note that this image will be planar, since the original instance from 3SAT is planar.

**Part I: CONSTRUCTION**

We construct a PLANAR 3SAT set of clauses into an equivalent theory that uses \( \Lambda(l, r) \) with the same construction we used in Theorem 10.3.

Replace each ternary clause

\[
\Lambda(l_1, r_1) \lor \Lambda(l_2, r_2) \lor \Lambda(l_3, r_3)
\]

by the set of lines and regions in Figure 10.4(a), where \( l_1, r_1, l_2, r_2, l_3, r_3 \) are parameters unique for each clause.

**Definition 10.25** A loop is a set \( I \) of lines and regions \( \{ l_1, r_1, ..., l_n, r_n \} \) such that for all \( 1 < i \leq n \) we have

1. 
\( I \models in(l_i, r_i) \land in(l_i, r_{i-1}) \land in(l_n, r_1) \)

2. 
\( I \models in(l_i, r_{i-1}) \land in(l_{i-1}, r_{i-1}) \land orphan(l_i) \land \neg \text{collinear}(l_i, l_{i-1}) \)

3. 
\( I \models in(l_n, r_n) \land in(l_1, r_n) \land orphan(l_n) \land \neg \text{collinear}(l_1, l_n) \)

An example of a loop is the image in Figure 10.5.

For each literal \( \Lambda(l_i, r_j) \) that appears in a binary and a ternary clause, \( r_j \) is a region in a loop. If \( \Lambda(l_i, r_j) \equiv \Lambda(l_k, r_m) \), then \( l_i, r_j \) are an even number of lines and regions away from \( l_k, r_m \) in the loop. If \( \Lambda(l_i, r_j) \equiv \neg \Lambda(l_k, r_m) \), then \( l_i, r_j \) are an odd number of lines and regions away from \( l_k, r_m \) in the loop. ⁵

**Part II: PROOF OF EQUIVALENCE**

---

⁵ Figure 10.11 is an example of the clause construct combined with a loop. The regions \( r_1, r_2, r_3 \) are the regions in the clause. The regions \( r_1, r_7, r_8, r_9, r_{10} \) are in one loop, the regions \( r_2, r_{11}, r_{12}, r_{14} \) are in the second loop, and \( r_3, r_4, r_5, r_6 \) are in the third loop.
Lemma 10.10 Consider the construct $I_1$ in Figure 10.4(a).

$$T_{cw} \cup I_1 \cup N.A.A \cup RDA \models \Lambda(l, r) \supset \Lambda(l_1, r_1) \lor \Lambda(l_2, r_2) \lor \Lambda(l_3, r_3)$$

Proof: Suppose

$$T_{cw} \cup I_1 \cup RDA \models \Lambda(l, r) \land \Delta(r, s)$$

then there are only three consistent poses for $s_1$, and the following pose clauses are satisfied by $T_{cw} \cup I_1 \cup RDA \cup N.A.A$:

$$\Lambda(l, r) \supset \Delta(r_1, s_1) \land \text{ontop}(s_1, s) \lor \Delta(r_2, s_2) \land \text{ontop}(s_2, s) \lor \Delta(r_3, s_3) \land \text{ontop}(s_3, s)$$

$$\Lambda(l_4, r_1) \supset \Delta(r, s) \land \text{ontop}(s, s_1)$$

$$\Lambda(l_5, r_2) \supset \Delta(r, s) \land \text{ontop}(s, s_2)$$

$$\Lambda(l_6, r_3) \supset \Delta(r, s) \land \text{ontop}(s, s_3)$$

By $RDA$ we have

$$\Lambda(l_1, r_1) \lor \Lambda(l_4, r_1)$$

$$\Lambda(l_2, r_2) \lor \Lambda(l_5, r_2)$$

$$\Lambda(l_3, r_3) \lor \Lambda(l_6, r_3)$$

Thus we get

$$T_{cw} \cup I_1 \cup RDA \models \Lambda(l, r) \supset \Lambda(l_1, r_1) \lor \Lambda(l_2, r_2) \lor \Lambda(l_3, r_3)$$

$\square$

Lemma 10.11 For all $i < n$ a loop satisfies

$$T_{cw} \cup I \cup RDA \cup N.A.A \models \Lambda(l_i, r_i) \equiv \neg \Lambda(l_{i-1}, r_{i-1})$$

$$T_{cw} \cup I \cup RDA \cup N.A.A \models \Lambda(l_{i-1}, r_i) \equiv \text{ontop}(s_{i-1}, s_i)$$

$$T_{cw} \cup I \cup RDA \cup N.A.A \models \Lambda(l_n, r_n) \equiv \neg \Lambda(l_1, r_n)$$

$$T_{cw} \cup I \cup RDA \cup N.A.A \models \Lambda(l_n, r_n) \equiv \text{ontop}(s_1, s_n)$$

Proof: By clauses of type (2) in $\Pi$, we have for all $i < n$,

$$\Lambda(l_i, r_{i-1}) \supset \neg \Lambda(l_{i-1}, r_{i-1})$$

$$\Lambda(l_n, r_n) \supset \neg \Lambda(l_1, r_n)$$

By $DTR$,

$$\neg \Lambda(l_{i-1}, r_{i-1}) \supset \text{ontop}(s_{i-1}, s_i)$$

$$\neg \Lambda(l_1, r_n) \supset \text{ontop}(s_1, s_n)$$

so that we have

$$\Lambda(l_i, r_{i-1}) \supset \text{ontop}(s_{i-1}, s_i)$$

$$\Lambda(l_n, r_n) \supset \text{ontop}(s_1, s_n)$$

For the other direction, by $RDA$ we have for all $i < n$,

$$\neg \Lambda(l_{i-1}, r_{i-1}) \supset \Lambda(l_i, r_{i-1})$$

$$\neg \Lambda(l_1, r_n) \supset \Lambda(l_n, r_n)$$
By DTR, we have
\[ \text{ontop}(s_{i-1}, s_i) \supset \neg A(l_{i-1}, r_{i-1}) \]
\[ \text{ontop}(s_1, s_n) \supset \neg A(l_1, r_n) \]
so that we get
\[ \text{ontop}(s_{i-1}, s_i) \supset A(l_i, r_{i-1}) \]
\[ \text{ontop}(s_1, s_n) \supset A(l_n, r_n) \]

Thus loops act as “communication” constructs by passing the value of variables between the different image constructs that simulate clauses. If the literal \(A(l_i, r_i)\) is assigned true in one clause, then in any clause containing that literal it must also be be also be assigned true; in any clause containing \(\neg A(l_i, r_i)\) the literal must be assigned false. The construction of the loop guarantees that these truth assignments are made consistently. Lines and regions an even distance from each other in the loop are given the same truth assignment while lines and regions an odd distance away from each other in the loop are given opposite truth assignments.

**Lemma 10.12** Loops and the construct in Figure 10.4 satisfy NCAO.

It is easy to see that the image can be built from the constructs in polynomial time and that the properties of the constructs ensure that \(T_{cw} \cup I \cup RDA \cup NAA\) is satisfiable iff the instance of PLANAR 3SAT is satisfiable. □

Figure 10.11 shows an example of the clause construct and associated loops. In this image, \(r_2, r_{11}, r_{12}, r_{14}\) are in one loop, \(r_3, r_4, r_5, r_6\) are in the second loop, and \(r_1, r_7, r_8, r_9, r_{10}\) are in the third loop. The arrows indicate the directions in which the loops continue.

### 10.4.4 Tractable Subclasses for Accidental Alignments

To find a tractable subclass, we need to remove the ambiguity of which regions depict surfaces and which are background. Thus we must look at images satisfying NAO. First we notice that if \(I \models NAO\), NAA is able to completely assign surface depiction literals to all regions in \(I\).

**Lemma 10.13** If \(I \models NAO\), then
\[ T_{cw} \cup I \cup RDA \cup NAA \models (\forall r) \text{region}(r) \supset [(\exists s) \Delta(r, s)] \]
\[ \exists l, l' \text{ in}(l, r) \land \text{parallel}(l, l') \lor \text{perp}(l, l') \land \text{between}(l, l', r) \]

**Proof:** Suppose
\[ T_{cw} \cup I \cup RDA \cup NAA \models \Delta(r_1, s_1) \]
By RDA there must be a line \(l_1\) and an edge \(e_1\) such that
\[ T_{cw} \cup I \cup RDA \cup NAA \models \text{in}(l_1, r_1) \land \Delta(l_1, e_1) \land \text{part}(e_1, s_1) \]
Since \(I \models NAO\), there are no orphans, so there must be another line \(l_2\) such that \(I \models \text{parallel}(l_1, l_2) \lor \text{perp}(l_1, l_2)\). By NAA, these two lines must satisfy
\[ T_{cw} \cup I \cup RDA \cup NAA \models \Delta(l_1, e_1) \land \Delta(l_2, e_2) \land (\text{parallel}(e_1, e_2) \lor \text{perp}(e_1, e_2)) \]
Figure 10.11: Clause construct combined with a loop.
Suppose

\[ I \models in(l_1, r_2) \land in(l_2, r_2) \land between(l_1, l_2, r_3) \land between(l_1, l_2, r_2) \land \lnot between(l_1, l_2, r_1) \]

Then by \( NAA \) we must have

\[ T_{cw} \cup I \cup RDA \cup NAA \models \Delta(l_1, e_1) \land \Delta(r_1, s_1) \supset \lnot part(e_1, s_1) \]

which contradicts the assumption above: therefore we must have \( I \models between(l_1, l_2, r_1) \).

Suppose

\[ T_{cw} \cup I \cup RDA \cup NAA \models in(l_1, r_1) \land (parallel(l_1, l_2) \lor perp(l_1, l_2)) \land between(l_1, l_2, r_1) \]

By \( NAA \) we must have

\[ T_{cw} \cup I \cup RDA \cup NAA \models \Delta(l_1, e_1) \land \Delta(l_2, e_2) \land part(e_1, s_1) \land part(e_2, s_1) \land (parallel(e_1, e_2) \lor perp(e_1, e_2)) \]

By the depiction axiom 9.5 of \( T_{cw} \) we get

\[ T_{cw} \cup I \cup RDA \cup NAA \models \Delta(r_1, s_1) \]

\( \square \)

Furthermore, if \( I \models NAO \) then \( NAA \) entails all region groupings.

**Lemma 10.14** If \( I \models NAO \), then

\[ T_{cw} \cup I \cup RDA \cup NAA \models (\forall r, r')(\exists s) \Delta(r, s) \land \Delta(r', s) \equiv (\exists l, l') in(l, r) \land in(l', r') \land (parallel(l, l') \lor perp(l, l')) \land between(l, l', r) \]

**Proof:** Suppose

\[ T_{cw} \cup I \cup RDA \cup NAA \models \Delta(r_1, s_1) \land \Delta(r_2, s_2) \]

Since \( I \models NAO \), by \( RDA \) there must be lines \( l_1, l_2 \) such that

\[ I \models in(l_1, r_1) \land in(l_2, r_2) \land (parallel(l_1, l_2) \lor perp(l_1, l_2)) \]

and by \( NAA \) there exist edges \( e_1, e_2 \) such that

\[ T_{cw} \cup I \cup RDA \cup NAA \models \Delta(l_1, e_1) \land part(e_1, s_1) \land \Delta(l_2, e_2) \land part(e_2, s_1) \]

\( RACA \) entails

\[ T_{cw} \cup I \cup RDA \cup NAA \models (parallel(e_1, e_2) \lor perp(e_1, e_2)) \]

By the depiction axiom 9.5 of \( T_{cw} \), we must also have \( I \models between(l_1, l_2, r_1) \).

Suppose

\[ T_{cw} \cup I \cup RDA \cup NAA \models in(l_1, r_1) \land in(l_2, r_1) \land (parallel(l_1, l_2) \lor perp(l_1, l_2)) \land between(l_1, l_2, r_1) \]

By \( NAA \) there must exist edges \( e_1, e_2 \) such that

\[ T_{cw} \cup I \cup RDA \cup NAA \models \Delta(l_1, e_1) \land \Delta(l_2, e_2) \land part(e_1, s_1) \land part(e_2, s_1) \]

\( RACA \) entails

\[ T_{cw} \cup I \cup RDA \cup NAA \models (parallel(e_1, e_2) \lor perp(e_1, e_2)) \]

\( RDA \) entails

\[ T_{cw} \cup I \cup RDA \cup NAA \models \Delta(r_1, s_1) \land \Delta(r_2, s_1) \]

\( \square \)
We now see that this assignment of surface depiction literals leads to a tractable algorithm.

**Theorem 10.12** Let I be the set of images satisfying \( \text{NAO} \). Then there is an \( O(|r|) \) algorithm to find a model of \( T_{cw} \cup I \cup \text{NAO} \cup \text{RDA} \) for some image \( I \in I \), where \( |r| \) is the number of regions in \( I \).

**Proof:** Since \( I \models \text{NAO} \), there do not exist lines \( l_i, l_j \) such that \( I \models \text{orphan}(l_i) \land \text{orphan}(l_j) \). By construction of \( \Pi_{\text{NAO}} \), all lines are assigned edge depiction literals. By the previous two lemmas, all regions are assigned surface depiction literals. Consistent poses are found for each surface using SELECT-POSE, since all region groupings are forced by \( \text{NAO} \).

The complexity of the algorithm is \( O(|r|) \) to assign the edge depiction literals and \( O(|r|) \) for SELECT-POSE.

**Corollary 10.2** If \( I \models \text{NAO} \) and \( T_{cw} \cup I \cup \text{RDA} \cup \text{NAO} \) is consistent, then it has a unique model.

### 10.4.5 A Computational Framework for \( \Pi_{\text{NAO}} \)

We will now consider the abductive non-accidental alignment assumption, \( \text{NAO} \), and construct a set of propositional clauses that is logically equivalent to it. This set of clauses will be used in the complexity proofs and in algorithms to construct models satisfying \( \text{NAO} \).

**Definition 10.26** \( \Pi_{\text{NAO}} \) is the following set of clauses:

1. For every lines \( l_i, l_j \) such that
   
   \[ I \models \text{in}(l_i, r_k) \land \text{in}(l_j, r_m) \land \text{between}(l_i, l_j, r_k) \land \text{between}(l_i, l_j, r_m) \land (\text{parallel}(l_i, l_j) \lor \text{perp}(l_i, l_j)) \]
   
   add the literals
   
   \[ \Lambda(l_i, r_k), \Lambda(l_j, r_m), \Lambda(l_i, r_m), \Lambda(l_j, r_k), \Delta(r_i, s_i), \Delta(r_j, s_i) \in \Pi_{\text{NAO}} \]

2. For all lines \( l_1, l_2 \) such that
   
   \[ I \models \text{in}(l_1, r_i) \land \text{in}(l_1, r_k) \land \text{in}(l_2, r_j) \land \text{in}(l_2, r_m) \land \text{collinear}(l_1, l_2) \land \text{orphan}(l_1) \land \text{orphan}(l_2) \]
   
   add the clauses
   
   \[ \Lambda(l_1, r_i) \lor \Lambda(l_2, r_m), \]
   
   \[ \Lambda(l_1, r_k) \lor \Lambda(l_2, r_j), \]
   
   \[ \Lambda(l_1, r_k) \lor \Delta(r_i, s_i), \Lambda(l_1, r_k) \lor \Delta(r_j, s_i), \]
   
   \[ \Lambda(l_1, r_i) \lor \Delta(r_k, s_k), \Lambda(l_1, r_i) \lor \Delta(r_m, s_k), \]
   
   \[ \Lambda(l_2, r_m) \lor \Delta(r_i, s_i), \Lambda(l_2, r_m) \lor \Delta(r_j, s_i), \]
   
   \[ \Lambda(l_2, r_j) \lor \Delta(r_k, s_k), \Lambda(l_2, r_j) \lor \Delta(r_m, s_k) \] \in \Pi_{\text{NAO}}

**Theorem 10.13** Let \( \Phi \) be a sentence in \( L(\Pi) \). Given a planar image \( I \), \( \Pi \cup \Pi_{\text{NAO}} \cup I \models \Phi \iff \Pi_{\text{NAO}} \cup I \cup \text{RDA} \cup \text{NAO} \models \Phi \).

**Proof:** This follows from the construction of \( \Pi_{\text{NAO}} \) and the equivalence \( \text{NAO} \equiv \text{NAO} \land \text{UDML} \). The clauses in \( \Pi_{\text{NAO}} \) are thus a subset of the clauses in \( \Pi_{\text{NAO}} \), simply excluding those clauses arising from \( \text{UDML} \). \( \square \)
We first notice that even if there are no ambiguous orphans, ANA runs into trouble. If \( I \models NAO \), then every line is assigned a depiction literal by ANA: also, non-orphans assign surface depiction literals to the region that they are in. However, not every region \( r \) has an orphan \( l \) such that \( I \models (l, r) \land (\text{parallel}(l, l') \land \text{perp}(l, l') \land \text{between}(l, l', r)) \): these regions are not assigned surface depiction literals by ANA. Thus we do not know whether such regions depict surfaces or not. For example, in Figure 10.2e, we cannot assign \( r_1 \) a surface depiction literal, since \( r_1 \) is not between two grouped lines. If such regions do depict surfaces, then by RDA one of the lines in the region must depict an edge in the surface: thus one of these lines must depict multiple edges, since by NAL all lines have been assigned depiction literals. We will now show that this ambiguity leads to intractability.

Note that if \( I \models NAO \), all depiction axioms have been satisfied since all lines have been assigned edge depiction literals. We thus need only consider the occlusion axioms when finding a model of \( T_{cw} \cup I \cup RDA \).

**Theorem 10.14** Let \( I = \{ I : I \models NAO \} \); then for \( I \in I \) it is NP-hard to find a model of \( T_{cw} \cup I \cup RDA \cup ANA \) or determine that a model does not exist.

**Proof:** We will show that the problem is NP-hard using a reduction from the NP-hard problem PLANAR 3SAT. Given an instance \( F \) of PLANAR 3SAT, we will construct an image \( I \) such that \( I \models NAL \) and such that \( F \) has a satisfying interpretation iff \( T_{cw} \cup I \cup ANA \cup RDA \) is consistent. Note that this image will be planar, since the original instance from 3SAT is planar.

We transform a PLANAR 3SAT set of clauses into an equivalent theory that uses \( \Lambda(l, r) \) with the same construction we used in Theorem 10.3.

Replace each ternary clause

\[
\Lambda(l_1, r_1) \lor \Lambda(l_2, r_2) \lor \Lambda(l_3, r_3)
\]

by the set of lines and regions in Figure 10.13, where \( l_1, r_1, l_2, r_2, l_3, r_3 \) are parameters unique for each clause.

**Definition 10.27** An occlusion loop is a set of lines and regions \( l_1, r_1, \ldots, l_n, r_n \) such that \( T_{cw} \cup I \cup ANA \cup RDA \) satisfies the following clauses:

1.

\[
I \models \text{in}(l_i, r_i) \land \text{in}(l_{n-i}, r_i)
\]

2.

\[
\Lambda(l_i, r_i) \supset \Lambda(l_{i-1}, r_{i-1}) \land \text{ontop}(s_{i-1}, s_i) \lor \Lambda(l_{n-i-1}, r_{i-1}) \land \text{ontop}(s_{i-1}, s_i)
\]

3.

\[
\Lambda(l_{n-i}, r_i) \supset \Lambda(l_{n-i-1}, r_{i-1}) \land \text{ontop}(s_{i-1}, s_i) \lor \Lambda(l_{i-1}, r_{i-1}) \land \text{ontop}(s_{i-1}, s_i)
\]

4.

\[
\Lambda(l_n, r_n) \supset \Lambda(l_1, r_1) \lor \text{ontop}(s_1, s_n) \lor \Lambda(l_{n-1}, r_1) \land \text{ontop}(s_1, s_n)
\]

5.

\[
\Lambda(l_{n-1}, r_1) \supset \Lambda(l_n, r_n) \lor \text{ontop}(s_n, s_1) \lor \Lambda(l_2n, r_n) \land \text{ontop}(s_n, s_1)
\]
An example of an occlusion loop is the image in Figure 10.12.
For each literal $\Lambda(l, r_j)$ that appears in a binary and a ternary clause, $r_j$ is a region in an occlusion loop. If $\Lambda(l, r_j) \equiv \Lambda(l_k, r_m)$, then $l_i, r_j$ are an even number of lines and regions away from $l_k, r_m$ in the occlusion loop. If $\Lambda(l, r_j) \equiv \neg \Lambda(l_k, r_m)$, then $l_i, r_j$ are an odd number of lines and regions away from $l_k, r_m$ in the occlusion loop.

Finally, for every clause of the form

$$\Lambda(l, r) \supset \Lambda(l_i, r_i) \lor \Lambda(l_j, r_j) \lor \Lambda(l_k, r_k)$$

where $r_i$ is in a loop with $n$ regions, $r_j$ is in a loop with $m$ regions and $r_k$ is in a loop with $p$ regions, add the literals $\Lambda(l, r) \land \text{ontop}(s, s)$ to every pose satisfying $\Lambda(l_{i-n}, r_i)$, add the literals $\Lambda(l, r) \land \text{ontop}(s, s)$ to every pose satisfying $\Lambda(l_{j-n}, r_j)$, and add the literals $\Lambda(l, r) \land \text{ontop}(s, s)$ to every pose satisfying $\Lambda(l_{k-n}, r_k)$.

In Figure 10.13, we thus have the following clauses:

$$\Lambda(l_1, r_1) \supset \Lambda(l_{10}, r_5) \land \text{ontop}(s_5, s_1)$$

$$\Lambda(l_4, r_1) \supset \Lambda(l_8, r_4) \land \text{ontop}(s_4, s_1) \land \Delta(r, s) \land \text{ontop}(s, s_1)$$

$$\Lambda(l_2, r_2) \supset \Lambda(l_{14}, r_9) \land \text{ontop}(s_9, s_2)$$

$$\Lambda(l_5, r_3) \supset \Lambda(l_{12}, r_8) \land \text{ontop}(s_8, s_2) \land \Delta(r, s) \land \text{ontop}(s, s_2)$$

$$\Lambda(l_3, r_3) \supset \Lambda(l_{18}, r_7) \land \text{ontop}(s_7, s_3)$$

$$\Lambda(l_6, r_3) \supset \Lambda(l_{16}, r_6) \land \text{ontop}(s_6, s_3) \land \Delta(r, s) \land \text{ontop}(s, s_3)$$

The regions $r_1, r_4, r_5$ are in one occlusion loop, $r_2, r_8, r_9$ are in the second occlusion loop, and $r_3, r_6, r_7$ are in the third occlusion loop.

**Part II: PROOF OF EQUIVALENCE**

**Lemma 10.15** Consider the construct $l_1$ in Figure 10.13. If

$$T_{cw} \cup I_1 \cup ANAA \cup RDA \models \Lambda(l, r) \land$$

$$\neg \text{ontop}(s_6, s_2) \land \neg \text{ontop}(s_2, s_6) \land \neg \text{ontop}(s_6, s_8) \land \neg \text{ontop}(s_8, s_6) \land$$

$$\neg \text{ontop}(s_9, s_{10}) \land \neg \text{ontop}(s_{10}, s_9) \land \neg \text{ontop}(s_6, s_2) \land \neg \text{ontop}(s_6, s_2)$$

then

$$T_{cw} \cup I_1 \cup ANAA \cup RDA \models$$

$$\Lambda(l_1, r_1) \lor \Lambda(l_2, r_2) \lor \Lambda(l_3, r_3)$$

**Proof:** We will use the partial assignment to constrain the possible poses for surfaces in the scene. Given the partial assignment of ontop literals, the following pose clauses are then satisfied by $T_{cw} \cup I_1 \cup ANAA \cup RDA$:

$$\Lambda(l, r) \supset \Delta(r_1, s_1) \lor \text{ontop}(s_1, s) \lor \Delta(r_2, s_2) \lor \text{ontop}(s_2, s) \lor \Delta(r_3, s_3) \lor \text{ontop}(s_3, s)$$

$$\Lambda(l_4, r_1) \supset \Delta(r, s) \lor \text{ontop}(s, s_1)$$

$$\Lambda(l_5, r_2) \supset \Delta(r, s) \lor \text{ontop}(s, s_2)$$

$$\Lambda(l_6, r_3) \supset \Delta(r, s) \lor \text{ontop}(s, s_3)$$

Thus we get

$$T_{cw} \cup I_1 \cup ANAA \cup RDA \models \Lambda(l, r) \supset \Lambda(l_1, r_1) \lor \text{ontop}(s_1, s) \lor \Lambda(l_2, r_2) \lor \text{ontop}(s_2, s) \lor \Lambda(l_3, r_3) \lor \text{ontop}(s_3, s)$$

Making the assignment $\Lambda(l, r)$ we get the clause

$$\Lambda(l_1, r_1) \lor \Lambda(l_2, r_2) \lor \Lambda(l_3, r_3)$$

$\square$
Figure 10.12: Occlusion loop.
Figure 10.13: Clause construct used in the proof for the Theorem 10.14.
Lemma 10.16 For all $i$, the occlusion loop $I = \{l_1, r_1, \ldots, l_n, r_n\}$ satisfies

$$T_{cw} \cup I \cup ANAA \cup RDA \models \Lambda(l_i, r_i) \equiv \Lambda(l_{i-1}, r_{i-1})$$

$$T_{cw} \cup I \cup ANAA \cup RDA \models \Lambda(l_i, r_i) \equiv \text{ontop}(s_{i-1}, s_i)$$

$$T_{cw} \cup I \cup ANAA \cup RDA \models \Lambda(l_i, r_i) \equiv \Lambda(l_n, r_n)$$

$$T_{cw} \cup I \cup ANAA \cup RDA \models \Lambda(l_i, r_i) \equiv \text{ontop}(s_1, s_n)$$

$$T_{cw} \cup I \cup ANAA \cup RDA \models \Lambda(l_i, r_i) \equiv \text{ontop}(s_{i-1}, s_i)$$

$$T_{cw} \cup I \cup ANAA \cup RDA \models \Lambda(l_i, r_i) \equiv \text{ontop}(s_1, s_{i-1})$$

$$T_{cw} \cup I \cup ANAA \cup RDA \models \Lambda(l_i, r_i) \equiv \text{ontop}(s_1, s_n)$$

$$T_{cw} \cup I \cup ANAA \cup RDA \models \Lambda(l_i, r_i) \equiv \text{ontop}(s_1, s_{i-1})$$

Proof: By conditions (2) and (3) of the construction, the pose clauses for two lines $l_i$ and $l_{n-i-1}$ are

$$\Lambda(l_i, r_i) \supset \text{ontop}(s_{i-1}, s_i) \lor \Lambda(l_{n-i-1}, r_{i-1}) \land \text{ontop}(s_{i-1}, s_i)$$

$$\Lambda(l_{n-i-1}, r_{i-1}) \supset \text{ontop}(s_{i-1}, s_i) \lor \Lambda(l_i, r_i) \land \text{ontop}(s_{i-1}, s_i)$$

Both of the poses for $\Lambda(l_{n-i-1}, r_{i-1})$ are inconsistent with the poses for $\Lambda(l_i, r_i)$, so that we have

$$\Lambda(l_i, r_i) \supset \neg \Lambda(l_{n-i-1}, r_{i-1})$$

By RDA we have

$$\Lambda(l_{i-1}, r_{i-1}) \lor \Lambda(l_{n-i-1}, r_{i-1})$$

which together give

$$\Lambda(l_i, r_i) \supset \Lambda(l_{i-1}, r_{i-1})$$

By DTR we have

$$\Lambda(l_i, r_i) \supset \text{ontop}(s_{i-1}, s_i)$$

The occlusion clause in $\Pi$ entails the clause

$$\Lambda(l_i, r_i) \supset \neg \text{ontop}(s_{i-1}, s_i)$$

Similarly, the lines $l_n, l_1$ in the loop have the pose clauses

$$\Lambda(l_n, r_n) \supset \text{ontop}(s_1, s_n) \lor \Lambda(l_{n-1}, r_1) \land \text{ontop}(s_1, s_n)$$

$$\Lambda(l_{n-1}, r_1) \supset \text{ontop}(s_1, s_n) \lor \Lambda(l_n, r_n) \land \text{ontop}(s_1, s_n)$$

Both of the poses for $\Lambda(l_{n-1}, r_1)$ are inconsistent with the poses for $\Lambda(l_n, r_n)$, so that we have

$$\Lambda(l_n, r_n) \supset \neg \Lambda(l_{n-1}, r_1)$$

By RDA we have

$$\Lambda(l_1, r_1) \lor \Lambda(l_{n-1}, r_1)$$

which together give

$$\Lambda(l_n, r_n) \supset \Lambda(l_1, r_1)$$

By DTR we have

$$\Lambda(l_n, r_n) \supset \text{ontop}(s_1, s_n)$$
The occlusion clause in $\Pi$ entails the clause
$$\Lambda(l_n, r_n) \supset \neg\text{ontop}(s_n, s_1)$$

Combining these two sets of clauses gives us
$$\Lambda(l_i, r_i) \supset \Lambda(l_n, r_n)$$
$$\Lambda(l_i, r_i) \supset \neg\Lambda(l_{2n}, r_n)$$
$$\Lambda(l_i, r_i) \supset \text{ontop}(s_n, s_1)$$
$$\Lambda(l_i, r_i) \supset \neg\text{ontop}(s_1, s_n)$$

and hence
$$\Lambda(l_{i-1}, r_{i-1}) \supset \Lambda(l_i, r_i)$$
$$\Lambda(l_n, r_n) \supset \Lambda(l_i, r_i)$$

The remainder of the equivalences easily follow from these. $\Box$

Thus occlusion loops act as "communication" constructs by passing the value of variables between the different image constructs that simulate clauses. If the literal $\Lambda(l_i, r_i)$ is assigned true in one clause, then in any clause containing that literal it must also be also be assigned true; in any clause containing $\neg\Lambda(l_i, r_i)$ the literal must be assigned false. The construction of the occlusion loop guarantees that these truth assignments are made consistently. Depiction literals for the lines $l_i$ in the occlusion loop are given the same truth assignment, while depiction literals for the lines $l_{i-1}$ are given opposite truth assignments.

**Lemma 10.17 Occlusion loops and the construct $l_i$ in Figure 10.13 satisfy NAO.**

It is easy to see that the image can be built from the constructs in polynomial time and that the properties of the constructs ensure that $T_{\text{cew}} \cup I \cup \text{ANAA} \cup \text{RDA}$ is satisfiable iff the instance of PLANAR 3SAT is satisfiable. $\Box$

### 10.4.6 Tractable Subclasses for Abductive Accidental Alignments

To find a tractable subclass, we need to remove the ambiguity of which regions depict surfaces and which are background. One way to do this is to require that all regions contain non-orphans that will assign surface depiction literals to the regions.

**Definition 10.28 The No Unassigned Regions (NUR) assumption is the sentence**

$$(\forall r) \text{region}(r) \supset (\exists l, l') \text{in}(l, r) \land (\text{parallel}(l, l') \lor \text{perp}(l, l')) \land \text{between}(l, l', r)$$

i.e., every region is between a pair of lines which are either parallel or perp to each other.

For example, the image in Figure 10.7e does not satisfy NUR, since none of the lines in region $r_1$ satisfy NUR. Similarly, there are images that NUR $\land \neg$NAO. However, we do have the following

**Proposition 10.3 If $I \models \text{NUR} \land \neg\text{NAO}$, then $T_{\text{cew}} \cup I \cup \text{RDA} \cup \text{ANAA}$ is inconsistent.**
Proof: Suppose NAO is falsified so that

\[ I \models orpan(l_1) \land in(l_1, r_1) \land in(l_1, r_2) \]

\( NUR \) entails the existence of lines \( l_2, l_3, l_4, l_5 \) such that

\[ I \models in(l_2, r_1) \land (\parallel(l_2, l_3) \lor \perp(l_2, l_3)) \land between(l_2, l_3, r_1) \]
\[ \land in(l_4, r_2) \land (\parallel(l_4, l_5) \lor \perp(l_4, l_5)) \land between(l_4, l_5, r_2) \]

\( ANAA \) entails

\[ T_{cw} \cup I \cup RDA \cup ANAA \models \Lambda(l_2, r_1) \land \Lambda(l_4, r_2) \]

Since

\[ I \models \neg(\parallel(l_1, l_2) \lor \perp(l_1, l_2) \lor \parallel(l_4, l_2) \lor \perp(l_4, l_2)) \]

we have the following clauses by construction of \( \Pi \):

\[ T_{cw} \cup I \cup RDA \cup ANAA \models (\Lambda(l_2, r_1) \supset \neg \Lambda(l_1, r_1)) \land (\Lambda(l_4, r_2) \supset \neg \Lambda(l_1, r_2) \]

contradicting the figure ground clause for \( l_1 \). \( \square \)

Using this assumption on the class of images, ANAA is able to completely assign surface depiction literals: this assignment leads to a tractable algorithm.

Lemma 10.18 If \( I \models NUR \land NAO \), then

\[ T_{cw} \cup I \cup RDA \cup ANAA \models (\forall r) \\text{region}(r) \supset (\exists s) \Delta(r, s) \]

Proof: Since \( I \models NUR \land NAO \), ANAA entails

\[ T_{cw} \cup I \cup RDA \cup ANAA \models (\forall r) \text{region}(r) \supset (\exists l, l', e, e', s) \text{in}(l, r) \land (\parallel(l, l') \lor \perp(l, l')) \land \text{between}(l, l', r) \land \Delta(l, e) \land \Delta(l', e') \land \text{part}(e, s) \land \text{part}(e', s) \]

The depiction axiom 9.6 of \( T_{cw} \) then entails

\[ T_{cw} \cup I \cup RDA \cup ANAA \models (\forall r) \text{region}(r) \supset (\exists s) \Delta(r, s) \]

\( \square \)

Notice the difference between this lemma and the analogous result for NAA. That lemma specified the necessary and sufficient conditions for a region to depict a surface; thus a region could be forced to be background by NAA. In the lemma for ANAA, all regions are forced to depict surfaces; ANAA is unable to force regions to be background. However, ANAA does entail all region groupings.

Lemma 10.19 If \( I \models NUR \land NAO \), then

\[ T_{cw} \cup I \cup RDA \cup ANAA \models (\forall r, r') (\exists s) \Delta(r, s) \land \Delta(r', s) \equiv \]

\[(\exists l, l') \text{in}(l, r) \land \text{in}(l', r') \land (\parallel(l, l') \lor \perp(l, l')) \land \text{between}(l, l', r) \]
Proof: Suppose 

\[ T_{cw} \cup I \cup RDA \cup \text{ANAA} \models \Delta(r_1, s_1) \land \Delta(r_2, s_2) \]

Since \( I \models NUR \land NAO \), by \( RDA \) there must exist lines \( l_1, l_2 \) such that 

\[ I \models \text{in}(l_1, r_1) \land \text{in}(l_2, r_2) \land (\text{parallel}(l_1, l_2) \lor \text{perp}(l_1, l_2)) \]

so that \( \text{ANAA} \) entails the existence of edges such that 

\[ T_{cw} \cup I \cup RDA \cup \text{ANAA} \models \Delta(l_1, e_1) \land \text{part}(e_1, s_1) \land \Delta(l_2, e_2) \land \text{part}(e_2, s_1) \]

By \( RACA \) and depiction axiom 9.5 in \( T_{cw} \), these two depicting lines must satisfy 

\[ T_{cw} \cup I \cup RDA \cup \text{ANAA} \models \text{parallel}(l_1, l_2) \lor \text{perp}(l_1, l_2) \]

By the depiction axiom 9.6 in \( T_{cw} \) we must also have 

\[ T_{cw} \cup I \cup RDA \cup \text{ANAA} \models \text{between}(l_1, l_2, r) \]

Now suppose 

\[ T_{cw} \cup I \cup RDA \cup \text{ANAA} \models \text{in}(l_1, r_1) \land \text{in}(l_2, r_2) \land (\text{parallel}(l_1, l_2) \lor \text{perp}(l_1, l_2)) \land \text{between}(l_1, l_2, r_1) \]

By \( \text{ANAA} \) there must exist a surface \( s_1 \) and edges \( e_1, e_2 \) such that 

\[ T_{cw} \cup I \cup RDA \cup \text{ANAA} \models \Delta(l_1, e_1) \land \Delta(l_2, e_2) \land \text{part}(e_1, s_1) \land \text{part}(e_2, s_1) \]

\( RDA \) and \( IDA \) entail 

\[ T_{cw} \cup I \cup RDA \cup \text{ANAA} \models \Delta(r_1, s_1) \land \Delta(r_2, s_1) \]

If \( I \models NAO \land \neg NUR \), then \( \text{ANAA} \) cannot completely assign edge depiction literals: this ambiguity led to the \( \text{NP} \)-hardness result earlier in this section. However, we now have the following:

**Lemma 10.20** If \( I \models NUR \land NAO \), then \( \text{ANAA} \) makes a complete edge depiction literal assignment to all lines in \( I \).

Proof: It is easy to see that if \( I \models NAO \), then \( \text{ANAA} \) assigns a positive depiction literal to all lines. We must show that each line that is in two regions is also assigned a negative depiction literal. Suppose 

\[ T_{cw} \cup I \cup RDA \cup \text{ANAA} \models \text{in}(l_1, r_1) \land \text{in}(l_1, r_2) \land \Lambda(l_1, r_1) \]

Since \( I \models NUR \), then there must also be lines \( l_2, l_3 \) such that 

\[ I \models \text{in}(l_2, r_2) \land \text{in}(l_3, r_2) \land \text{between}(l_2, l_3, r_2) \land (\text{parallel}(l_2, l_3) \lor \text{perp}(l_2, l_3)) \]

so that 

\[ T_{cw} \cup I \cup RDA \cup \text{ANAA} \models \Lambda(l_2, r_2) \land \Lambda(l_3, r_2) \]

If 

\[ I \models \neg (\text{parallel}(l_1, l_2) \lor \text{parallel}(l_1, l_3) \lor \text{perp}(l_1, l_2) \lor \text{perp}(l_1, l_3)) \]

then by construction of \( \Pi \) we have the clauses 

\[ T_{cw} \cup I \cup RDA \cup \text{ANAA} \models (\Lambda(l_2, r_2) \lor \neg \Lambda(l_1, r_2)) \land (\Lambda(l_3, r_2) \lor \neg \Lambda(l_1, r_2)) \]

and thus 

\[ T_{cw} \cup I \cup RDA \cup \text{ANAA} \models \neg \Lambda(l_1, r_2) \]

\( \Box \)
Theorem 10.15 If \( I \models NUR \land NAO \), then there is an \( O(|r|) \) algorithm to find a model of \( T_{cw} \cup I \cup ANAA \cup RDA \) where \(|r|\) is the number of regions in \( I \).

Proof: Since \( I \models NAO \), there do not exist lines \( l_i, l_j \) such that \( I \models \text{orphan}(l_i) \land \text{orphan}(l_j) \). By the previous three lemmas, all regions are assigned surface depiction literals and all lines are completely assigned edge depiction literals. Consistent poses are found for each surface using SELECT-POSE, since all region and background groupings are forced.

The complexity of the algorithm is \( O(|r|) \) to assign the edge depiction literals and \( O(|r|) \) for SELECT-POSE. \( \square \)

Corollary 10.3 If \( I \models NUR \land NAO \) and \( T_{cw} \cup I \cup RDA \cup ANAA \) is consistent, then it has a unique model.

Proof: As demonstrated in the previous theorem, all lines are assigned edge depiction literals and all regions are assigned surface depiction literals. This assignment is obviously unique. \( \square \)

Rather than define a subclass of images to achieve tractability, another approach is to introduce a new depiction assumption.

Definition 10.29 The Background Depiction Assumption (BDA) is the sentence

\[ (\forall r) \text{ region}(r) \supset ((\exists s) \Delta(r, s) \equiv (\exists l, l') \text{ in}(l, r) \land (\text{parallel}(l, l') \lor \text{perp}(l, l')) \land \text{between}(l, l', r)) \]

i.e., a region depicts a surface iff it is between two lines that are parallel or perp to each other.

Using this assumption we get the following result:

Theorem 10.16 If \( I \models NAO \), there is an \( O(|r|) \) algorithm to find a model of \( T_{cw} \cup I \cup RDA \cup ANAA \cup BDA \), where \(|r|\) is the number of regions in \( I \).

Proof: Since \( I \models NAO \), all lines are assigned edge depiction literals; BDA will then assign surface depiction literals to all regions. We thus have a complete assignment of depiction literals. Consistent poses are found for each surface using SELECT-POSE, since all region and background groupings are forced by ANAA.

The complexity of the algorithm is \( O(|r|) \) to assign the edge depiction literals and \( O(|r|) \) for SELECT-POSE. \( \square \)

10.5 Region-restricted Accidental Alignments

The next step is to investigate assumptions weaker than NAA (so that they allow accidental alignments) but which still allow the existence of polynomial algorithms. For example, there may be exceptions to NAA, such as proximity, that may block grouping in some cases. Informally, [Jacobs 88] uses such heuristics; in this section we formalize this idea by introducing the non-accidental alignment assumptions of the previous sections, but restrict them to single regions.
10.5.1 Definitions for Region-based Restrictions

Using the grouping assumptions $\text{N.AA}$ and $\text{A.N.A.A}$ from the previous section, we can restrict the notion of accidental alignment to single regions as follows:

**Definition 10.30** Region-restricted NAA ($\text{RNAA}$) is the conjunction of the following sentences

$$(\forall l, l', e, r) \text{ collinear}(l, l') \land \text{in}(l, r) \land \text{in}(l', r) \land \Delta(l, e) \supset \Delta(l', e)$$

$$(\forall l, l', e, e', s) \text{ parallel}(l, l') \land \text{in}(l, r) \land \text{in}(l', r) \land \Delta(l, e) \land \Delta(l', e') \land \text{part}(e, s) \supset \text{part}(e', s)$$

$$(\forall l, l', e, e', s) \text{ perp}(l, l') \land \text{in}(l, r) \land \text{in}(l', r) \land \Delta(l, e) \land \Delta(l', e') \land \text{part}(e, s) \supset \text{part}(e', s)$$

**Definition 10.31** Region-restricted A.N.A.A ($\text{RANAA}$) is the conjunction of the following sentences

$$(\forall l, l', r) \text{ collinear}(l, l') \land \text{in}(l, r) \land \text{in}(l', r) \supset (\exists e, s) \Delta(l, e) \land \Delta(l', e) \land \text{part}(e, s)$$

$$(\forall l, l', r) \text{ parallel}(l, l') \land \text{in}(l, r) \land \text{in}(l', r) \supset (\exists e, e', s) \Delta(l, e) \land \Delta(l', e') \land \text{part}(e, s) \land \text{part}(e', s)$$

$$(\forall l, l', r) \text{ perp}(l, l') \land \text{in}(l, r) \land \text{in}(l', r) \supset (\exists e, e', s) \Delta(l, e) \land \Delta(l', e') \land \text{part}(e, s) \land \text{part}(e', s)$$

**Proposition 10.4** For any image $I,$

$$T_{cw} \cup I \models \text{N.AA} \supset \text{RNAA}$$

$$T_{cw} \cup I \models \text{A.N.A.A} \supset \text{RANAA}$$

This proposition easily follows from the definitions. Thus an image may have no region-restricted accidental alignments, but still have accidental alignments.

**Example:** In Figure 10.14(a), there is an accidental alignment, since $I \models \text{parallel}(l_1, l_2)$ but these lines cannot depict edges in the same surface. However, there are no region-restricted accidental alignments. Similarly, in Figure 10.14(b), there are accidental alignments, since

$$I \models \text{perp}(l_1, l_2) \land \text{parallel}(l_1, l_3) \land \text{perp}(l_2, l_3)$$

yet all three lines cannot depict edges in the same surface. There are no region-restricted accidental alignments since the image has a model satisfying

$$\mathcal{M} \models \Delta(l_1, e_1) \land \Delta(l_2, e_2) \land \Delta(l_3, e_3) \land \Delta(l_4, e_4) \land \Delta(l_5, e_5) \land \Delta(l_6, e_6) \land \Delta(l_7, e_7) \land \Delta(l_8, e_8) \land \Delta(r_1, s_1) \land \Delta(r_2, s_2) \land \Delta(r_3, s_3) \land \text{part}(e_1, s_1) \land \text{part}(e_2, s_2) \land \text{part}(e_3, s_3) \land \text{part}(e_4, s_2) \land \text{part}(e_5, s_2) \land \text{part}(e_6, s_1) \land \text{part}(e_7, s_3) \land \text{part}(e_8, s_3)$$

The image in Figure 10.14(c) has both accidental and region-restricted accidental alignments, since we have

$$T_{cw} \cup I \models -\Theta(l_1, r_1) \land \Theta(l_2, r_1) \land \Theta(l_3, r_1)$$

$\square$
Figure 10.14: Example images for region-restricted accidental alignments.
Definition 10.32 Given an image $I$, a region-restricted orphan $l$ is a line in $I$ that satisfies

$$(\forall l) \text{region\_orphan}(l) \equiv \text{line}(l) \land \neg (\exists l', r \in (l, r) \land \text{in}(l', r) \land (\parallel(l, l') \lor \perp(l, l')))$$

i.e., a line is a region-restricted orphan iff there does not exist any other line in the region which is parallel or perp with it.

Definition 10.33 An ambiguous region-restricted orphan $l$ is a line that satisfies

$$(\forall l) \text{aro}(l) \equiv \text{region\_orphan}(l) \land \text{ambiguous}(l) \land \neg (\exists l', l'', r \in (l, r) \land \text{in}(l', r) \land \text{in}(l'', r) \land (\parallel(l', l'') \lor \perp(l', l'')))$$

i.e., a line is an ambiguous region-restricted orphan iff it is a region-restricted orphan, it is ambiguous, and there does not exist any other lines in a region containing it which are parallel or perp to each other.

The following is an easy consequence of the definitions:

Proposition 10.5 For any image $I$,

$$T_{cw} \cup I \models (\forall l) \text{orphan}(l) \supset \text{region\_orphan}(l)$$

We can therefore have lines which are not orphans, but which are still region orphans.

As with the relationship between $NAA$ and $ANAA$, we can decompose $RNAA$ into $RANAA$ and a unique depiction assumption:

Definition 10.34 The Uniquely Depicting Region-restricted Orphan (UDRO) assumption is the sentence

$$(\forall l, e, s, r, e') \text{region\_orphan}(l) \land \text{in}(l, r) \land \Delta(l, e) \land \text{part}(e, s) \land \Delta(l, e') \supset e = e'$$

i.e., region orphans depict unique edges.

Theorem 10.17 For any image $I$,

$$T_{cw} \cup I \models RNAA \equiv RANAA \land UDRO$$

Proof: It is easy to see that

$$T_{cw} \cup I \models RNAA \supset RANAA$$

Suppose then that an image $I$ violates $UDRO$, so that there exists lines $l_1, l_2$, a region $r_1$, and edges $e_1, e_2$ such that

$$T_{cw} \cup I \models (\parallel(l_1, l_2) \lor \perp(l_1, l_2) \lor \text{collinear}(l_1, l_2)) \land \text{in}(l_1, r_1) \land \text{in}(l_2, r_1) \land \Delta(l_1, e_1) \land \Delta(l_1, e_2) \land e_1 \neq e_2$$

However, we then have

$$T_{cw} \cup I \models \neg (\exists s) (\text{part}(e_1, s) \land \text{part}(e_2, s))$$

violating $RNAA$.

To prove the other direction of the equivalence we will show

$$T_{cw} \cup I \models RANAA \land \neg RNAA \supset \neg UDRO$$

Thus, suppose we have a region-restricted accidental alignment which satisfies $RANAA$:

$$T_{cw} \cup I \models (\parallel(l_1, l_2) \lor \perp(l_1, l_2)) \land \text{in}(l_1, r_1) \land \text{in}(l_2, r_1) \land \Delta(l_1, e_1) \land \Delta(l_2, e_2)$$

$$\land \text{part}(e_1, s_1) \land \text{part}(e_2, s_1) \land \neg \text{part}(e_3, s_1)$$

It is easy to see that this violates $UDRO$. □
Thus \( \text{R.A.N.A.A.} \) allows some region restricted non-orphans to depict multiple edges, just as \( \text{A.V.A.A.} \) allows some non-orphans to depict multiple edges. As with \( \text{U.D.M.L.} \), the assumption \( \text{U.D.R.O.} \) applies only to multiple lines depicting edges in a surface, including non-region restricted orphans and collinear region-restricted orphans. \( \text{U.D.R.O.} \) forces these lines to depict unique edges. On the other hand, orphans for which there is no collinear line can depict multiple edges, as in Figure 10.6(d), where \( I \models \text{aro}(l_i) \).

### 10.5.2 Classes of Images for Region-Restricted Alignments

As we did earlier with ambiguous lines and orphans, we can use the notion of region-restricted orphan to define image classes that will be used in the complexity proofs.

**Definition 10.35** The No Ambiguous Region-restricted Orphan (\( \text{NARO} \)) assumption is the sentence

\[
(\forall l) \text{line}(l) \supset \neg \text{aro}(l)
\]

Note that \( \text{NARO} \) entails

\[
(\forall l) \text{line}(l) \supset (\exists l', r) \text{in}(l', r) \wedge (\text{parallel}(l, l') \vee \text{perp}(l, l')) \wedge \neg \text{ambiguous}(l)
\]

i.e., for every line in a region, there exists another line in the same region which is either parallel or perpendicular to it.

The images in Figure 10.15(a) and Figure 10.15(b) satisfy \( \text{NARO} \).

**Definition 10.36** The No Collinear Ambiguous Region-restricted Orphan (\( \text{NCARO} \)) assumption is the sentence

\[
(\forall l, l') \text{aro}(l) \wedge \text{aro}(l') \supset \neg \text{collinear}(l, l')
\]

i.e., there are no collinear ambiguous region-restricted orphans in the image.

In Figure 10.15c, there are three ambiguous region-restricted orphans \( l_2, l_3, l_4 \); since none of these lines are collinear with any other lines, this image satisfies \( \text{NCARO.} \) Also note that none of these lines are orphans, since

\[ I \models \text{parallel}(l_1, l_2) \wedge \text{parallel}(l_3, l_4) \]

**Definition 10.37** The Isolated Ambiguous Region-restricted Orphan (\( \text{IARO} \)) assumption is the sentence

\[
(\forall l, l', r) \text{aro}(l) \wedge \text{aro}(l') \wedge \text{in}(l, r) \wedge l \neq l' \supset \neg \text{in}(l', r)
\]

i.e., there is at most one ambiguous region-restricted orphan in a region.

In Figure 10.15c, there are two regions \( r_1, r_2 \) such that

\[ I \models \text{in}(l_1, r_1) \wedge \text{in}(l_2, r_1) \wedge \text{in}(l_2, r_2) \wedge \text{in}(l_3, r_2) \wedge \text{ambiguous}(l_1) \wedge \text{ambiguous}(l_2) \]

so that this image does not satisfy \( \text{IARO.} \) In Figure 10.15d, there are only two ambiguous lines \( l_1, l_2 \), and they are the only ambiguous lines in their regions; thus this image satisfies \( \text{IARO.} \) However, since

\[ I \models \text{collinear}(l_1, l_2) \]

this image does not satisfy \( \text{NCARO.} \) Again, note that none of the region-restricted orphans are orphans.

The following relationships among these classes of images are easy consequences of the definitions:
Figure 10.15: Classes of images related to region-restricted accidental alignments.
Proposition 10.6 For any image \( I \),
\[
T_{cw} \cup I \models NARO \supset IARO
\]
\[
T_{cw} \cup I \models NARO \supset NAO
\]

The relationships among the image classes is shown in Figure 10.16. The arrows represent entailment between the two sentences defining the image classes.

One drawback of region-restricted grouping assumptions is that regions groupings will no longer be entailed. There may be lines that are region-restricted orphans but that are not orphans: such lines will not be grouped by RNAA or RANAA. Another approach to use with these grouping assumptions is to consider RNAA and RANAA as giving higher priority to region-restricted alignments: we could then use assumptions such as NAA or ANAA to group the non-orphans in the image that are in different regions. This would allow us to construct models for images that had certain accidental alignments. On the other hand, there is no inherent preference in RNAA or RANAA alone for any models satisfying region groupings or nonaccidental alignments of nonorphans.

10.5.3 Computational Framework for RNAA

As with the previous assumptions, we will define a propositional theory \( \Pi \cup \Pi_{RNAA} \), which is logically equivalent to RNAA.
Definition 10.38 $\Pi_{\text{RNAA}}$ is the following set of clauses:

1. For all lines $l_i, l_j$ such that

$$I \models \text{in}(l_i, r_k) \land \text{in}(l_j, r_k) \land \text{in}(l_i, r_m) \land \text{in}(l_j, r_n) \land \text{between}(l_i, l_j, r_k)$$

and $I \models \text{parallel}(l_i, l_j)$ or $I \models \text{perp}(l_i, l_j)$, or $I \models \text{collinear}(l_i, l_j)$, then add the literals

$$\Lambda(l_i, r_k), \Lambda(l_j, r_k), \Lambda(r_k, s_k) \in \Pi_{\text{RNAA}}$$

2. If

$$I \models \text{in}(l_1, r_{i1}) \land \text{in}(l_1, r_{k1}) \land \text{in}(l_2, r_{i1}) \land (\text{parallel}(l_1, l_2) \lor \text{perp}(l_1, l_2) \lor \text{collinear}(l_1, l_2))$$

introduce the clause

$$\neg \Lambda(l_1, r_{i1}) \lor \neg \Lambda(l_1, r_{k1}) \in \Pi_{\text{RNAA}}$$

Theorem 10.18 Let $\Phi$ be a sentence in $\mathcal{L}(\Pi)$. Given a planar image $I$, $\Pi \cup \Pi_{\text{RNAA}} \cup I \models \Phi$ iff $T_{\text{cw}} \cup I \cup \text{RDA} \cup \Pi_{\text{RNAA}} \models \Phi$.

Proof: In the previous chapter we proved that the clauses in $\Pi$ are equivalent to $T_{\text{cw}} \cup \text{RDA} \cup I$ for some image $I$. We simply need to show that N.A.A is equivalent to the clauses in $\Pi_{\text{RNAA}}$ for some image $I$.

1. If

$$I \models \text{in}(l_1, r_{k1}) \land \text{in}(l_2, r_{k1}) \land (\text{parallel}(l_1, l_2) \lor \text{perp}(l_1, l_2))$$

then RNAA reduces the depiction axiom 9.5 in $T_{\text{cw}}$ to

$$(\forall l, l', r, a) \text{in}(l, r) \land \text{in}(l', r) \land \text{between}(l, l', r) \land \theta(a, l, l') \land a < \pi \supset \Lambda(l, r) \land \Lambda(l', r)$$

If

$$I \models \text{in}(l_1, r_{k1}) \land \text{in}(l_2, r_{k1}) \land \text{collinear}(l_1, l_2)$$

then RNAA reduces the depiction axiom 9.6 in $T_{\text{cw}}$ to

$$(\forall l, l', r, a) \text{in}(l, r) \land \text{in}(l', r) \land \text{between}(l, l', r) \supset \Lambda(l, r) \land \Lambda(l', r)$$

2. Suppose

$$I \models \text{in}(l_1, r_{i1}) \land \text{in}(l_1, r_{k1}) \land (\text{parallel}(l_1, l_2) \lor \text{perp}(l_1, l_2) \lor \text{collinear}(l_1, l_2))$$

By R.A.A we must have

$T_{\text{cw}} \cup \text{RDA} \cup I \models \Delta(l_1, e_1) \land \Delta(l_2, e_2) \land \text{part}(e_1, s_1) \land \text{part}(e_2, s_1)$

By I.D.A we must have

$T_{\text{cw}} \cup \text{RDA} \cup I \models \Delta(r_1, s_1) \land \Delta(r_2, s_2)$

so that in this case $U\text{DRO}$ is equivalent to

$$(\forall l, l') \text{in}(l, r) \land \text{in}(l', r') \supset (\Lambda(l, r) \supset \neg \Lambda(l, r'))$$

Notice that no ambiguity arises from $\text{collinear}$ as it did with N.A.A; since accidental alignments are restricted to unique regions, there is no figure ground ambiguity with a collinear region-restricted orphan. However, region-restricted orphans do lead to problems, even when they are not orphans.
Theorem 10.19 Let \( \mathcal{I} = \{ I : I \models \text{NAO} \} \); then for \( I \in \mathcal{I} \) it is NP-hard to find a model of \( T_{cw} \cup I \cup \text{RDA} \cup \text{RNAA} \) or determine that a model does not exist.

Proof: We will show that the problem is NP-hard using a reduction from PLANAR 3SAT. Given an instance \( F \) of PLANAR 3SAT, we will construct an image \( I \) such that \( I \models \text{NAO} \) and such that \( F \) has a satisfying interpretation if \( T_{cw} \cup I \cup \text{RDA} \cup \text{RNAA} \) is consistent. Note that this image will be planar, since the original instance from 3SAT is planar.

The proof will be similar to the NP-hardness proof for \( T_{cw} \cup I \cup \text{RDA} \cup \text{RNAA} \) when \( I \models \text{NAO} \), except we will use region-restricted orphans that are not orphans.

Part I: THE CONSTRUCTION

We construct transform a PLANAR 3SAT set of clauses into an equivalent theory that uses \( \Lambda(l, r) \) with the same construction we used in Theorem 10.3.

Replace each ternary clause
\[-\Lambda(l_1, r_1) \lor \Lambda(l_2, r_2) \lor \Lambda(l_3, r_3)\]
by the set of lines and regions in Figure 10.17, where \( l_1, r_1, l_2, r_2, l_3, r_3 \) are parameters unique for each clause.

Definition 10.39 A region-restricted loop is a set of regions \( \{ r_1, \ldots, r_n \} \) such that for all \( i \leq n \) we have

1. \( I \models \text{in}(l_i, r_i) \land \text{in}(l_i, r_{i-1}) \)

2. \( I \models \text{in}(l_i, r_i) \land \text{in}(l_{i-1}, r_{i-1}) \land \text{aro}(l_i) \)

3. \( I \models \text{in}(l_n, r_n) \land \text{in}(l_1, r_1) \)

4. \( I \models \text{in}(l_n, r_n) \land \text{in}(l_1, r_n) \land \text{aro}(l_n) \)

An example of a region-restricted loop is in Figure 10.18.

For each literal \( \Lambda(l_i, r_j) \) that appears in binary and ternary clause, \( r_j \) is a region in a region-restricted loop.

Part II: PROOF OF EQUIVALENCE

Lemma 10.21 The construct \( I_1 \) in Figure 10.4 satisfies:

\[ T_{cw} \cup I_1 \cup \text{RNAA} \cup \text{RDA} \models \Lambda(l_1, r_1) \lor \Lambda(l_2, r_2) \lor \Lambda(l_3, r_3) \]

Proof: This follows from the proof for Theorem 10.11, which uses the same construct. \( \square \)

Lemma 10.22 For all \( i \leq n \) a a region-restricted loop satisfies

\[ T_{cw} \cup I \cup \text{RDA} \cup \text{RNAA} \models \Lambda(l_i, r_i) \equiv \neg \Lambda(l_{i-1}, r_{i-1}) \]

\[ T_{cw} \cup I \cup \text{RDA} \cup \text{RNAA} \models \Lambda(l_i, r_i) \equiv \text{ontop}(s_{i-1}, s_i) \]

\[ T_{cw} \cup I \cup \text{RDA} \cup \text{RNAA} \models \Lambda(l_n, r_n) \equiv \neg \Lambda(l_1, r_n) \]

\[ T_{cw} \cup I \cup \text{RDA} \cup \text{RNAA} \models \Lambda(l_n, r_n) \equiv \text{ontop}(s_1, s_n) \]
Figure 10.17: Clause construct used in the proof of Theorem 10.19.
Figure 10.18: Region-restricted loop.
Proof: By clauses of type (2) in $\Pi$, we have for all $i < n$,

$$
\Lambda(l_i, r_{i-1}) \supset \neg \Lambda(l_{i-1}, r_{i-1}) \\
\Lambda(l_n, r_n) \supset \neg \Lambda(l_1, r_n)
$$

By DTR,

$$
\neg \Lambda(l_{i-1}, r_{i-1}) \supset \text{ontop}(s_{i-1}, s_i) \\
\neg \Lambda(l_1, r_n) \supset \text{ontop}(s_1, s_n)
$$

so that we have

$$
\Lambda(l_i, r_{i-1}) \supset \text{ontop}(s_{i-1}, s_i) \\
\Lambda(l_n, r_n) \supset \text{ontop}(s_1, s_n)
$$

For the other direction, by RDA we have for all $i < n$,

$$
\neg \Lambda(l_{i-1}, r_{i-1}) \supset \Lambda(l_i, r_{i-1}) \\
\neg \Lambda(l_1, r_n) \supset \Lambda(l_n, r_n)
$$

By DTR, we have

$$
\text{ontop}(s_{i-1}, s_i) \supset \neg \Lambda(l_{i-1}, r_{i-1}) \\
\text{ontop}(s_1, s_n) \supset \neg \Lambda(l_1, r_n)
$$

so that we get

$$
\text{ontop}(s_{i-1}, s_i) \supset \Lambda(l_i, r_{i-1}) \\
\text{ontop}(s_1, s_n) \supset \Lambda(l_n, r_n)
$$

As with previous proofs, region-restricted loops act as "communication" constructs by passing the value of variables between the different image constructs that simulate clauses. If the literal $\Lambda(l_i, r_i)$ is assigned true in one clause, then in any clause containing that literal it must also be assigned true; in any clause containing $\neg \Lambda(l_i, r_i)$ the literal must be assigned false. The construction of the loop guarantees that these truth assignments are made consistently. Lines and regions an even distance from each other in the loop are given the same truth assignment while lines and regions an odd distance away from each other in the loop are given opposite truth assignments.

Lemma 10.23 Region-restricted loops and the construct $I_1$ in Figure 10.17 satisfy NAO.

It is easy to see that the image can be built from the constructs in polynomial time and that the properties of the constructs ensure that $T_{cu} \cup I \cup RDA \cup RNAA$ is satisfiable iff the instance of PLANAR 3SAT is satisfiable. $\square$

10.5.4 Tractable Subclasses for RNAA

As with the other grouping assumptions, we must therefore move to a more restricted class of images to achieve a tractable algorithm. It is insufficient to require that there be no orphans in the image. Since our grouping assumption has been restricted to regions, we must enforce the condition that there be no region-restricted orphans. The following lemmas show that by doing so, RNAA is able to completely assign depiction literals to the lines and regions in the image; this sets the stage for a polynomial algorithm.
Lemma 10.24 If \( I \models NARO \), then
\[
T_{cw} \cup I \cup RDA \cup RNAA \models (\forall r) \text{region}(r) \supset (\exists s) \Delta(r, s)
\]

Proof: Suppose Since \( I \models NARO \), by \( RDA \) there must exist lines \( l_1, l_2 \) such that
\[
I \models \text{in}(l_1, r_1) \land \text{in}(l_2, r_1) \land (\text{parallel}(l_1, l_2) \lor \text{perp}(l_1, l_2))
\]
so that \( RNAA \) entails the existence of edges such that
\[
T_{cw} \cup I \cup RDA \cup RNAA \models \Delta(l_1, e_1) \land \text{part}(e_1, s_1) \land \Delta(l_2, e_2) \land \text{part}(e_2, s_1)
\]
By \( RACA \) and depiction axiom 9.5 in \( T_{cw} \), these two depicting lines must satisfy
\[
T_{cw} \cup I \cup RDA \cup RNAA \models \text{parallel}(l_1, l_2) \lor \text{perp}(l_1, l_2))
\]
Suppose
\[
I \models \text{in}(l_1, r_2) \land \text{in}(l_2, r_2) \land \text{between}(l_1, l_2, r_2) \land \neg \text{between}(l_1, l_2, r_1)
\]
Then by \( U DRO \) we must have
\[
T_{cw} \cup I \cup RDA \cup RNAA \models \Delta(l_1, e_1) \land \Delta(r_1, s_1) \supset \neg \text{part}(e_1, s_1)
\]
which contradicts our hypothesis. We must therefore have \( I \models \text{between}(l_1, l_2, r_1) \).
Now suppose
\[
T_{cw} \cup I \cup RDA \cup RNAA \models \text{in}(l_1, r_1) \land \text{in}(l_2, r_1) \land (\text{parallel}(l_1, l_2) \lor \text{perp}(l_1, l_2)) \land \text{between}(l_1, l_2, r_1)
\]
By \( RNAA \) there must exist a surface \( s_1 \) such that
\[
T_{cw} \cup I \cup RDA \cup RNAA \models \Delta(l_1, e_1) \land \Delta(l_2, e_2) \land \text{part}(e_1, s_1) \land \text{part}(e_2, s_1)
\]
\( RDA \) and \( IDA \) entail
\[
T_{cw} \cup I \cup RDA \cup RNAA \models \Delta(r_1, s_1)
\]
\( \Box \)

Notice that although all regions are assigned a surface depiction literal, \( RNAA \) does not entail any region groupings, since it is restricted to lines in the same region.

Lemma 10.25 If \( I \models NARO \), then all lines are given a complete edge depiction literal assignment by \( RNAA \).

Proof: This follows from the construction of \( \Pi_{RNAA} \). \( \Box \)

Theorem 10.20 If \( I \models NARO \), then there is an \( O(|r|^2) \) algorithm to find a model of \( T_{cw} \cup I \cup RDA \cup RNAA \).

Proof: Since \( I \models NARO \), there do not exist lines \( l_i, l_j \) such that \( I \models \text{aro}(l_i) \land \text{aro}(l_j) \). By construction of \( \Pi_{RNAA} \), all lines are assigned edge depiction literals. By the previous lemma, all regions are assigned surface depiction literals. Consistent poses are found for each surface using \( \text{POSE-CONSISTENCY} \).

The complexity of the algorithm is \( O(|r|) \) to assign the edge depiction literals and \( O(|r|^2) \) for \( \text{POSE-CONSISTENCY} \). \( \Box \)

Corollary 10.4 If \( I \models NARO \) and \( T_{cw} \cup I \cup RDA \cup RNAA \) is consistent, then it has a unique model.
10.5.5 Computational Framework for RANAA

We will now use the propositional theory $\Pi \cup \Pi_{\text{RANAA}}$ for computing models of $T_{cw} \cup I \cup RDA \cup\text{RANAA}$.

Definition 10.40 $\Pi_{\text{RANAA}}$ is the following set of literals:

For all lines $l_i, l_j$ such that $I \models \text{in}(l_i, r_k) \land \text{in}(l_j, r_k)$ and $I \models \text{parallel}(l_i, l_j)$ or $I \models \text{perp}(l_i, l_j)$, or $I \not\models \text{collinear}(l_i, l_j)$, add the literals

$$\lambda(l_i, r_k), \lambda(l_j, r_k), \Delta(r_k, s_k) \in \Pi_{\text{RANAA}}$$

Theorem 10.21 Let $\Phi$ be a sentence in $L(\Pi)$. Given an image $I$, $\Pi \cup \Pi_{\text{RANAA}} \cup I \models \Phi$ iff $T_{cw} \cup I \cup RDA \cup\text{RANAA} \models \Phi$.

Proof: This follows from the proof for $\Pi_{\text{RANAA}}$ since $\text{RANAA} \equiv \text{RANAA} \land \text{UDRO}$.

Theorem 10.22 Let $\mathcal{I} = \{ I : I \models \text{NARO} \}$; then for $I \in \mathcal{I}$ it is NP-hard to find a model of $T_{cw} \cup I \cup RDA \cup\text{RANAA}$ or determine that a model does not exist.

Proof: This follows from the NP-hardness proof for $T_{cw} \cup I \cup RDA$, since both the clause construct $I_1$ and the occlusion loop $I_2$ satisfy NARO, and $\Pi \models \Pi_{\text{RANAA}}$ for these images.

As with ANAA, the problem with RANAA is that there is ambiguity in the assignment of depiction literals: we do not know which regions depict surfaces and which are background, and there is no complete assignment of edge depiction literals to the lines in the image. Previously, we introduced an additional constraint on the class of images; we now introduce a region-restricted version of that constraint.

Definition 10.41 The Region-restricted No Unassigned Regions (RNUR) assumption is the following sentence

$$(\forall r) \text{region}(r) \supset (\exists l, l') \text{in}(l, r) \land \text{in}(l', r) \land (\text{parallel}(l, l') \lor \text{perp}(l, l') \lor \text{collinear}(l, l'))$$

i.e., every region contains a pair of lines which are either parallel, perp, or collinear.

Lemma 10.26 If $I \models \text{RNUR} \land \text{NARO}$, then

$$T_{cw} \cup I \cup RDA \cup\text{RANAA} \models (\forall r) \text{region}(r) \supset (\exists s) \Delta(r, s)$$

Proof: Since $I \models \text{RNUR} \land \text{NARO}$, then RANAA entails

$$T_{cw} \cup I \cup RDA \cup\text{RANAA} \models (\forall r) \text{region}(r) \supset (\exists l, l', e, e', s) \text{in}(l, r) \land \text{in}(l', r) \land (\text{parallel}(l, l') \lor \text{perp}(l, l'))$$

$$\land \text{between}(l, l', r) \land \Delta(l, e) \land \Delta(l', e') \land \text{part}(e, s) \land \text{part}(e', s)$$

The depiction axiom 9.6 of $T_{cw}$ then entails

$$T_{cw} \cup I \cup RDA \cup\text{ANAA} \models (\forall r)(\exists s) \Delta(r, s)$$

Notice that although all regions are assigned a surface depiction literal, RANAA does not entail any region groupings, since it is restricted to lines in the same region.

Theorem 10.23 If $I \models \text{RNUR} \land \text{NARO}$, then there is an $O(|r|^2)$ algorithm to find a model of $T_{cw} \cup I \cup RDA \cup\text{RANAA}$ for some image $I \in \mathcal{I}$.
Proof: Since \( I \models RNU R \), there do not exist lines \( l_i, l_j \) such that \( I \models aro(l_i) \land aro(l_j) \). By construction of \( \Pi_{RANAA} \), all lines are assigned edge depiction literals. By the previous lemma, all regions are assigned surface depiction literals. Consistent poses are found for each surface using POSE-CONSISTENCY.

The complexity of the algorithm is \( O(|r|) \) to assign the edge depiction literals and \( O(|r|^2) \) for POSE-CONSISTENCY. \( \square \)

As we did with \( ANAA \), instead of defining a subclass of images, we can introduce a new depiction assumption to achieve tractability.

Definition 10.42 The Region-restricted Background Depiction Assumption (RBDA) is the sentence

\[(\forall r) \text{ region}(r) \supset ((\exists s) \Delta(r, s) \equiv (\exists l, l') \text{ in}(l, r) \land \text{in}(l', r) \land (\text{parallel}(l, l') \lor \text{perp}(l, l') \land \text{collinear}(l, l') \land \text{between}(l, l', r)))\]

i.e., a region depicts a surface iff it contains two lines which are either parallel, perp, or collinear.

Using this assumption we get the following result:

Theorem 10.24 If \( I \models NARO \), there is an \( O(|r|^2) \) algorithm to find a model of \( T_{cw} \cup I \cup RDA \cup RANAA \cup RBDA \), where \( |r| \) is the number of regions in \( I \).

Proof: Since \( I \models NARO \), all lines are assigned edge depiction literals; RBDA will then assign surface depiction literals to all regions. We thus have a complete assignment of depiction literals. Consistent poses are found for each surface using POSE-CONSISTENCY.

The complexity of the algorithm is \( O(|r|) \) to assign the edge depiction literals and \( O(|r|^2) \) for POSE-CONSISTENCY. \( \square \)
Figure 10.19: Complexity Results for Grouping Assumptions
10.6 Summary

In this chapter, we have presented a formalization of the notion of grouping and explored the role it plays in the complexity of the task of object recognition. We have shown that the general problem of finding the model of an image for the various grouping assumptions is NP-hard. However, we have identified sentences that define classes of images for which the task of finding a model for the grouping assumptions is tractable.

The complexity results of this chapter are summarized in Figure 10.19. The dashed lines indicate the "computational cliff" for theories consisting of different grouping and depiction assumptions. If an image is in a class below the dashed line for a particular theory, then there exists a polynomial algorithm to find a model of the theory; if the image is in a class above the dashed line, then in general it is NP-hard to find a model of the theory. For example, if $I \models NARO$, then the task of finding a model of $T_{cw} \cup I \cup RAD \cup RANAA$ is NP-hard, but the task of finding a model of $T_{cw} \cup I \cup RAD \cup NAA$ is polynomial. If $I \models NAO$, then the task of finding a model of $T_{cw} \cup I \cup RAD \cup NAA$ is tractable, but if $I \models IAO$ and $I \not\models NAO$, then the task of finding a model of $T_{cw} \cup I \cup RAD \cup NAA$ is NP-hard.

Intuitively, for a given grouping assumption, there is a threshold of ambiguity which is related to the degree of occlusion in the scene. If too much of a surface is occluded, then there will be insufficient cues for the grouping assumption to disambiguate the assignment of depiction literals to lines and regions.

We can also use this as a "meta-heuristic" for discovering new accidental alignment assumptions – once we identify the property of the image or scene which allows ambiguity, we can introduce a new assumption which provides a unique figure/ground assignment for the specific class of images or scenes.
Chapter 11

Summary and Future Work

11.1 Summary

The main objective of this thesis has been the axiomatization of some of the intuitions for image understanding. In the first set of results, we have introduced eight theories collectively referred to as the CardWorld Theories:

- $T_{scene}$: axiomatizes the basic ontology of scenes within CardWorld, consisting of surfaces, edges, and points. (Chapter 3)

- $T_{scene} \cup T_{kernel}$: extends $T_{scene}$ with the basic ontology of images (consisting of regions, lines, and pixels), as well as the basic intuitions about the depiction relation between scene objects and image objects. (Chapter 3)

- $T_{fg} \cup T_{scene}$: extends $T_{scene}$ by axiomatizing intuitions about figure and ground for 2D polygonal surfaces, and provides a basic theory of shape. (Chapter 4)

- $T_{\Delta} \cup T_{fg} \cup T_{kernel} \cup T_{scene}$: extends the previous theories by axiomatizing the depiction of the figure/ground relations in $T_{fg}$. (Chapter 5)

- $T_{interior} \cup T_{fg} \cup T_{scene}$: extends the intuitions about figure and ground to define the notions of the interior of a surface (including any holes that the surface may have). (Chapter 6)

- $T_{\Delta}^{\Delta} \cup T_{interior} \cup T_{\Delta} \cup T_{fg} \cup T_{kernel} \cup T_{scene}$: extends the previous one by axiomatizing the depiction of the interior of a surface and the relationship to properties of regions in an image. (Chapter 6)

- $T_{obscur} \cup T_{interior} \cup T_{fg} \cup T_{scene}$: introduces the notion of obscuring surfaces within a scene. (Chapter 7)

- $T_{occ} \cup T_{obscur} \cup T_{\Delta} \cup T_{interior} \cup T_{\Delta} \cup T_{fg} \cup T_{kernel} \cup T_{scene}$: this theory, which is the union of all of the CardWorld modules, axiomatizes the relationship between the notion of obscuring surfaces and nondepiction of obscured scene objects. (Chapter 7)
These theories provide an axiomatization of images and depiction for 2D polygonal surfaces in scenes with occlusion and noise (errors in edge detection). For each theory, we have provided a characterization of a class of structures, and shown that the theory is satisfied by these structures, and that all models of the theory are structures in the class.

In addition, we have explored various assumptions about shape and errors in edge detection which can be used to restrict the models of the CardWorld Theories to be finite. In particular, we showed that the assumption of no errors in edge detection is equivalent to a set of ten independent assumptions about depiction of surfaces.

The second set of major results explored the complexity of finding models of the CardWorld theories. Chapter 10 provided the basic theorem showing that the problem is in general NP-hard. However, we discovered that the fundamental reason for the intractability arises from ambiguity in the depiction of edges and surfaces. We then introduced a set of depiction assumptions related to accidental alignment, which determine which depicted edges belong to the same surface. If the theory entails these assumptions, then the complexity of finding a model becomes polynomial.

These results also allow an assumption-based approach to complexity results. Rather than specifying special tractable cases using syntactic criteria (such as 2SAT or Horn clauses), we can introduce sentences in the language of CardWorld and consider the complexity of finding a model of the extended theory.

11.2 Future Work

There are two major directions in which to extend the work of this thesis. First, we can define new axiomatic theories:

- Consider theories in which there are new classes of scene objects, such as colour, surface markings, shadows, and texture. This would require the relaxation of axiom 3.16 in $T_{scene}$, although other axioms would be unchanged (since they explicitly refer to points, edges, or surfaces).

- Explore axiomatizations of three-dimensional scene objects.

- Identify and axiomatize new accidental alignment assumptions, particularly in theories which allow errors in edge detection. This could possibly lead to a better account of perceptual organization.

- Axiomatizations that incorporate intuitions about granularity: The problem of scale has emerged as a fundamental source of difficulty in the qualitative description of images, because the events we perceive and find meaningful vary enormously in size and extent. An adequate representation must provide descriptions at a variety of physical scales, both in the image and in the scene.

- Apply the methodology used in this thesis to extend the Mapsee axiomatization. There are intuitions within the Mapsee domain which have analogues to figure/ground assignments, e.g. determining which side of a shoreline is the lake and which side is solid ground.
The second direction is to extend the complexity results for the existing axiomatic theories:

- All of the assumptions used in this thesis were either nonaccidental depiction assumptions or restrictions on the classes of images. Another possibility is to introduce scene assumptions: for example, we can consider scenes in which all of the surfaces are layered, in that there is a partial ordering over the surfaces with respect to ontop. All of the NP-hardness results in this thesis correspond to non-layered scenes. Motivation for considering layered scenes also arises from other concerns. Work in computer graphics has often used the Painter's Algorithm ([Newell et al 72], [Fiume 89]) to construct images from a scene. A necessary condition for this algorithm and the Painter's Algorithm is that the set of surfaces in the scene are ordered: in a non-layered scene, this is not the case, and the Painter's Algorithm will not work. As a point of trivia, the cover of the *Journal of Computational Geometry* depicts a non-layered scene.

- Similarly, we can impose additional depiction assumptions independent of nonaccidental alignment. For example, we can consider the complexity of finding a model given that we make the assumption that all lines depict unique edges.

- We can relax the assumption that all surfaces in the scene are convex, and allow nonconvex surfaces. This has deep ramifications, since many of the tractability results depended on all edges being convex, since this allowed us to infer the figure/ground assignment from the alignment of the depicting lines alone. This is no longer possible for nonconvex surfaces: for example, in an L-shaped surface, there are two pairs of perpendicular edges which have different figure/ground assignments.

- We can relax the shape closure assumption to allow a library of multiple shapes. [Grimson 90] provides a complexity analysis of his interpretation tree algorithm for a library of shapes, and it would be interesting to formally characterize the complexity of this task. In particular, this leads to heuristics which are based on the ratio of scene clutter to object size: the CardWorld theories can be used to specify these heuristics as assumptions within the language of $T_{cw}$.

- One of the major assumptions made in the complexity results was $NLEA$ – there are no errors in edge detection. Given the decomposition of this assumption into ten independent depiction assumptions, there is a vast array of theories which can be explored in which we weaken these assumptions and allow errors in edge detection. For example, we can consider the complexity of finding models of theories which contain silhouettes (and hence violated $URDA$).

- We can represent accidental alignments as defaults rather than first-order sentences. It is important to note that we will be constructing models of $T_{cw} \cup I \cup RDA \cup NAA$ and $T_{cw} \cup I \cup RDA \cup ANAA$. Thus, if a model has accidental alignments, then it will be inconsistent with these grouping
assumptions. An alternative approach would be to construct models of images that have accidental alignments, but in which we prefer those models that minimize the accidentals.
Appendix A

Hilbert’s Axiomatization of Geometry

The axioms in this section are taken from [Hilbert 71] and [Greenberg 93]. In the original presentation of his theory, Hilbert only provided English statements for the axioms; he did not provide a set of first-order sentences in standard notation, or even formally specify his language, as we have done in this thesis.

The language of Hilbert’s axiomatization of geometry is a second-order language, with the nonlogical lexicon:

\[ \mathcal{L}_\text{Hilbert} = \{ \text{point}(p), \text{line}(l), \text{on}(p,l), \text{ordered}(x,y,z), \text{congruent}(x,y,z,w), \text{segment}(x,y), \text{angle}(x,y,z), \text{parallel}(x,y) \} \]

- The unary predicate \text{point}(p) denotes the relation defining the class of points.
- The unary predicate \text{line}(l) denotes the relation defining the class of lines.
- The binary predicate \text{on}(p,l) denotes the incidence relation between points and lines.
- The ternary predicate \text{ordered}(x,y,z) denotes the relation whose intended interpretation is that the point \( y \) is between point \( z \) and point \( x \).
- The four-place predicate \text{congruent} denotes the relation whose intended interpretation is that the segment \( \overline{xy} \) is congruent (equal in magnitude) to the segment \( \overline{zw} \).
- The intended interpretation of the ternary predicate \text{segment}(x,z,y) is that \( z \) is a point on a line between points \( x \) and \( y \).
- The intended interpretation of the ternary predicate \text{angle}(x,y,z) is that the two segments \( \overline{x,y} \) and \( \overline{y,z} \) form an angle.
- The binary predicate \text{parallel}(x,y) denotes that two lines \( x,y \) are parallel, and do not intersect each other.
A.1 The Axioms

A.1.1 Incidence Axioms

Axiom 1 For every point \( P \) and for every point \( Q \) not equal to \( P \) there exists a unique line \( l \) incident with \( P \) and \( Q \).

Axiom 2 For every line \( l \) there exists at least two distinct points that are incident with \( l \).

Axiom 3 There exist three distinct points with the property that no line is incident with all three of them.

A.1.2 Order Axioms

Axiom 4 If \( B \) is between \( A \) and \( C \), then \( A, B, \) and \( C \) are three distinct points all lying on the same line, and \( B \) is between \( C \) and \( A \).

Axiom 5 Given any two distinct points \( B \) and \( D \), there exist points \( A,C \), and \( E \) lying on the line \( 
 \) such that \( B \) is between \( A \) and \( D \), \( C \) is between \( B \) and \( D \), and \( D \) is between \( B \) and \( E \).

Axiom 6 If \( A, B, \) and \( C \) are three distinct points lying on the same line, then one and only one of the points is between the other two.

A.1.3 Congruence Axioms

Axiom 7 If \( A \) and \( B \) are distinct points and if \( A' \) is any point, then for each ray \( r \) emanating from \( A' \) there is a unique point \( B' \) on \( r \) such that \( B' \neq A' \) and \( AB \cong A'B' \).

Axiom 8 If \( AB \cong CD \), and \( AB \cong EF \), then \( CD \cong EF \). Moreover, every segment is congruent to itself.

Axiom 9 If \( B \) is between \( A \) and \( C \), \( B' \) is between \( A' \) and \( C' \), \( AB \cong A'B', BC \cong B'C', AC \cong A'C' \).

Axiom 10 Given an angle \( \angle BAC \) and given any ray \( \overline{A'B'} \) emanating from a point \( A' \), then there is a unique ray \( \overline{A'C'} \) on a given side of the line \( \overline{A'B'} \) such that \( \angle B'A'C' \cong \angle BAC \).

Axiom 11 If \( \angle A \cong \angle B \) and \( \angle A \cong \angle C \), then \( \angle B \cong \angle C \). Moreover, every angle is congruent to itself.

Axiom 12 If two sides and the included angle of one triangle are congruent to two sides and the included angle of another triangle, then the two triangles are congruent.

A.1.4 Parallelism Axiom

Axiom 13 For every line \( l \) and every point \( P \) not lying on \( l \) there is at most one line \( m \) through \( P \) such that \( m \) is parallel to \( l \).

A.1.5 Completeness Axiom

Axiom 14 Suppose that the set \( \{l\} \) of all points on a line \( l \) is the disjoint union \( \Sigma_1 \cup \Sigma_2 \) of two nonempty subsets such that no point of either subset is between points of the other. Then there exists a unique point \( O \) on \( l \) such that one of the subsets is equal to a ray of \( l \) with vertex \( O \) and the other subset is equal to the complement.
A.2 Additional Relations

Definition A.1 Suppose we are given a structure $\mathcal{M} \in \mathcal{M}_{\text{polygon}}$, and let $\mathcal{H}$ be the model of Hilbert’s axiomatization of geometry.

An edge $e_1 \in \mathcal{M}$ is bounded by two other edges $e_2, e_3$ iff one of the following conditions holds:

1. If $(a_1, e_2, e_1) \not\in \Theta$, $(a_2, e_2, e_3) \not\in \Theta$, and $a_1 < \pi$, $a_2 < \pi$, then there exist points $x_1, x_2, x_3 \in H$ such that

   $$(x_1, h(e_1)) \in \text{on}, (x_2, h(e_2)) \in \text{on}, (x_3, h(e_3)) \in \text{on}$$

   and

   $$(x_1, x_2, x_3) \in \text{ordered}$$

2. If $a_1, (e_2, e_1) \in \Theta$, $(a_2, e_2, e_3) \in \Theta^M$, and $a_1 < \pi$, $a_2 > \pi$, then there do not exist points $x_1, x_2, x_3 \in H$ such that

   $$(x_1, h(e_1)) \in \text{on}, (x_2, h(e_2)) \in \text{on}, (x_3, h(e_3)) \in \text{on}$$

   and

   $$(x_1, x_2, x_3) \in \text{ordered}$$

3. If $(e_2, e_1) \in \text{parallel}$, then there exist points $x_1, x_2, x_3 \in H$ such that

   $$(x_1, h(e_1)) \in \text{on}, (x_2, h(e_2)) \in \text{on}, (x_3, h(e_3)) \in \text{on}$$

   and

   $$(x_1, x_2, x_3) \in \text{ordered}$$

4. If $(a_1, e_2, e_1) \in \Theta$, $(a_2, e_2, e_3) \in \Theta$, and $a_1 = \pi$, $a_2 = \pi$, then there exist points $x_1, x_2, x_3 \in H$ such that

   $$(x_1, h(e_1)) \in \text{on}, (x_2, h(e_2)) \in \text{on}, (x_3, h(e_3)) \in \text{on}$$

   and

   $$(x_1, x_2, x_3) \in \text{ordered}$$
Appendix A

Index of Assumptions

All assumptions labelled by acronyms are presented in alphabetical order. The number in parentheses is the page on which the assumption is axiomatized.

ANAA : Abductive Non-Accidental Alignment assumption (345)

BDA : Background Depiction Assumption (373)

BREA : Bounded Region Existence Assumption (274)

CCA : Connected Contour Assumption (248)

CFGA : Convex Figure/Ground Assumption (294)

CSA : Convex Surfaces Assumption (294)

DKA : Depicted Kernel Assumption (45)

DKRE : Depicted Kernel Region Existence Assumption (266)

DTR : Depicted Trichotomy assumption (257)

HREA : Hole Region Existence Assumption (259)

IAL : Isolated Ambiguous Lines assumption (335)

IAO : Isolated Ambiguous Orphans assumption (349)

IARO : Isolated Ambiguous Region-restricted Orphans assumption (377)

IDA : Depiction Assumption for the in relation (263)

IDKA : Depicting Kernel Assumption (49)

INDA : Inclusion Depiction Assumption (252)

MBRE : Multiple Background Region Existence assumption (271)
NAA : Non-Accidental Alignment assumption (345)
NAL : No Ambiguous Lines assumption (333)
NAO : No Ambiguous Orphans assumption (348)
NARO : No Ambiguous Region-restricted Orphans assumption (377)
NCAL : No Collinear Ambiguous Lines assumption (335)
NCAO : No Collinear Ambiguous Orphans assumption (349)
NCARO : No Collinear Ambiguous Region-restricted Orphans assumption (377)
NLEA : No Lacunate Errors Assumption (240)
NPAL : No Pairs of Ambiguous Lines assumption (335)
NPAO : No Pairs of Ambiguous Orphans assumption (349)
NSEA : No Spurious Errors Assumption (243)
NUR : No Unassigned Regions assumption (370)
OSDA : Occluded Surface Depiction Assumption (280)
PSCA : Parametrized Shape Closure Assumption (293)
RACA : Rectangular Alignment Closure Assumption (292)
RANAA : Region-restricted Abductive No Accidental Alignment assumption (374)
RDA : Region Depiction Assumption (303)
REA : Region Existence Assumption (266)
RFGC : Rectangular Figure/Ground Closure assumption (293)
RNAA : Region-restricted No Accidental Alignment assumption (374)
RNUR : Region-restricted No Unassigned Regions assumption (386)
RSCA : Rectangular Shape Closure Assumption (292)
SDKA : Strong Depicted Kernel Assumption (267)
UDML : Uniquely Depicting Multiple Lines assumption (347)
UDRO : Uniquely Depicting Region-restricted Orphans assumption (376)
URDA : Unique Region Depiction Assumption (251)

WCCA : Weak Connected Contour Assumption (248)

WDTR : Weak Depicted Trichotomy assumption (258)

WREA : Weak Region Existence Assumption (268)
Bibliography


