SORTAH:
System Optimal Re-Routing Transit Assignment Heuristic

by

Landy Ling Cheung

A thesis submitted in conformity with the requirements for the degree of Master of Applied Science
Civil Engineering
University of Toronto

© Copyright by Landy Ling Cheung 2015
Abstract

Traditionally, transit networks are modelled to follow the user equilibrium principle where passengers take a path that would benefit them the most. This does not directly address the system perspective and results in inferior performance compared to the pure system optimal model for congested networks. This research presents a new framework for the transit assignment problem to address the system perspective of an overcrowded transit network while integrating individual travel needs. It proposes a novel approach called System Optimal Re-Routing Transit Assignment Heuristic (SORTAH) that minimizes the overall congestion effects on a transit network. SORTAH leads to a better performing solution than that of a user equilibrium model, while simultaneously guaranteeing superior fairness compared to the pure system optimal model. A proof-of-concept prototype is implemented to compare network performance of the user equilibrium solution and the SORTAH solution. A large-scale real-world application is conducted on the Toronto Transit Commission (TTC) network.
Acknowledgments

This thesis and research would not have been possible without the help and assistance of many people.

First and foremost, I would like to express my sincere gratitude to my supervisor, Dr. Amer S. Shalaby, for the advice, guidance, and expertise he provided me throughout this research process. I cannot thank him enough for his moral support, mentoring and patience throughout my graduate studies. His supervision has set for me an example that I would like to follow in my future career.

I would also like to thank many colleagues at the transportation group at the University of Toronto. I gained insight and different perspectives of my research through our debates and discussions.

Last but not least, I want to express my heartfelt thanks to my family and Zhan Hong Liau for their overwhelming support and love. I would not have completed this thesis and research without their patience and continuous encouragement.
# Table of Contents

Acknowledgments ........................................................................................................ iii

Table of Contents .......................................................................................................... iv

List of Tables .................................................................................................................... ix

List of Figures .................................................................................................................. x

Chapter 1 Introduction .................................................................................................... 1
  1 Chapter Overview ........................................................................................................ 1
  2 Problem Statement ..................................................................................................... 1
  3 Motivation .................................................................................................................. 3
  4 Scope and Objectives ................................................................................................ 8
  5 Thesis Structure ......................................................................................................... 10

Chapter 2 Literature Review .......................................................................................... 12
  1 Chapter Overview ........................................................................................................ 12
  2 Early Developments .................................................................................................. 12
    2.1 Shortest-Path Algorithm ..................................................................................... 13
    2.2 Strategy and Hyperpath ..................................................................................... 14
  3 User Equilibrium Paradigm ....................................................................................... 16
    3.1 Classical Static User Equilibrium ...................................................................... 16
    3.2 Stochastic User Equilibrium ............................................................................. 18
    3.3 Dynamic User Equilibrium .............................................................................. 20
  4 System Optimal Paradigm ........................................................................................ 24
5 Complex Problem Solving Techniques ................................................................. 27
  5.1 Mathematical Programming ........................................................................... 27
    5.1.1 Method of Successive Averages (MSA) ..................................................... 27
    5.1.2 Frank-Wolfe (FW) .................................................................................. 28
    5.1.3 Gradient Projection (GP) ........................................................................ 28
  5.2 Constraint Programming ................................................................................ 29
  5.3 Heuristics ....................................................................................................... 32

6 Discussion ........................................................................................................... 33

Chapter 3 Transit Performance Measurement System ............................................. 35
  1 Chapter Overview ............................................................................................... 35
  2 Purpose of Performance Measurement Systems .................................................. 35
  3 Stakeholders of Transit Networks ...................................................................... 36
  4 Types of Performance Measures ........................................................................ 38
    4.1 Route-Level Performance Measures .............................................................. 38
      4.1.1 Passenger Loads ..................................................................................... 39
      4.1.2 Reliability ............................................................................................... 40
      4.1.3 Travel Time ............................................................................................ 41
    4.2 System-Level Performance Measures ........................................................... 43
  5 The SORTAH Performance Measurement System .............................................. 44
    5.1 Goals and Objectives .................................................................................... 44
    5.2 The Performance Measures .......................................................................... 44
      5.2.1 Load Factor ............................................................................................. 45
      5.2.2 Measure of Unfairness .......................................................................... 51
      5.2.3 Measures of Overcrowding ................................................................. 52
      5.2.4 Perceived Travel Time Measures ......................................................... 53
Chapter 4 SORTAH: System Optimal Re-routing Transit Assignment Heuristic............ 57

1 Chapter Overview ........................................................................................................... 57
2 Basic Assumptions ....................................................................................................... 57
3 Modelling Framework ................................................................................................. 58
   3.1 Network Representation ......................................................................................... 59
   3.2 Passenger’s Path Set ............................................................................................. 60
   3.3 Path Allocation Model ........................................................................................... 65
      3.3.1 Re-Routing Criteria ....................................................................................... 66
      3.3.2 Travel Time Adjustment ............................................................................... 68
      3.3.3 Re-Routing Evaluation ................................................................................... 70
4 The Heuristic Procedure ............................................................................................... 70
5 Characteristics and Limitations ................................................................................... 73
   5.1 Characteristics ......................................................................................................... 73
   5.2 Limitations ............................................................................................................... 74
6 Application Opportunities ........................................................................................... 76
   6.1 Policies and Strategies ........................................................................................... 76
   6.2 ITS Applications ..................................................................................................... 76
   6.3 Transit Route Design ............................................................................................. 77

Chapter 5 Prototype Implementation ................................................................................ 78

1 Chapter Overview ......................................................................................................... 78
2 Prototype Description ................................................................................................... 78
   2.1 MILATRAS ............................................................................................................. 78
   2.2 Variations of the SORTAH Prototype ................................................................... 80
3 Scope of Analysis ........................................................................................................... 84
4 Results and Discussion .................................................................................................. 85
  4.1 Crowding Levels ........................................................................................................ 85
  4.2 Total Travel Time ...................................................................................................... 89
  4.3 Execution Time .......................................................................................................... 91
5 Conclusion .................................................................................................................... 93

Chapter 6 Case Study – The Toronto Transit Commission (TTC) Application.............. 97

1 Chapter Overview ........................................................................................................ 97
2 Application Background ............................................................................................. 97
  2.1 The Transportation Tomorrow Survey (TTS)......................................................... 97
  2.2 The Toronto Transit Commission (TTC) System............................................... 99
3 Scope of Analysis ......................................................................................................... 101
4 Data Requirements ....................................................................................................... 101
  4.1 MILATRAS Outputs ............................................................................................... 102
    4.1.1 Demand ........................................................................................................... 102
    4.1.2 Supply ........................................................................................................... 104
  4.2 Re-Routing Criteria ................................................................................................. 104
5 Results and Discussion ................................................................................................. 105
6 Conclusion .................................................................................................................... 109

Chapter 7 Conclusion ....................................................................................................... 110

1 Chapter Overview ........................................................................................................ 110
2 Key Findings ................................................................................................................ 110
  2.1 SORTAH Prototypes ............................................................................................... 110
    2.1.1 Crowding Levels ............................................................................................ 111
    2.1.2 Total Travel Time ......................................................................................... 111
List of Tables

Table 1. Levels of Service at Different Load Factors (TRB, 2013) ........................................ 39
Table 2. Level of Service at Different Average Standing Passenger Space (TRB, 2013) .......... 40
Table 3. Perceived Travel Time Factors for Siting and Standing Passengers (TRB, 2013) .... 42
Table 4. Level of Service at Different Transit-Auto Travel Time Ratios (TRB, 2013) .......... 43
Table 5. Load Factor LOS A to LOS C Standards for All Transit Modes (TRB, 2013) ......... 47
Table 6. Typical Seat and Crush Capacities of Bus and Rail Transit Modes ...................... 48
Table 7. Load Factor Level of Service Standard - Bus Mode ............................................. 49
Table 8. Load Factor Level of Service Standard - Light Rail Mode ..................................... 50
Table 9. Load Factor Level of Service Standard - Heavy Rail Mode .................................. 50
Table 10. Relative Travel Time Component Factors (TRB, 2013) ...................................... 56
Table 11. Transit Unit Standing Times for Various Transit Vehicles (Vuchic, 2005) .......... 69
Table 12. Measure of Route Overcrowding ........................................................................ 85
Table 13. Percentage Change in Route Overcrowding ..................................................... 87
Table 14. Percentage Change in Segment Overcrowding ............................................... 88
Table 15. Percentage Change in Total Travel Time .......................................................... 90
Table 16. Summary of Percentage Change in Crowding Levels and Total Travel Time ...... 93
Table 17. Comparison of the Three Travel Time Criteria Cases ....................................... 96
Table 18. GTA and Surrounding Regions’ Zone-Coding for Year 2001 (Joint Program in Transportation, 2003a) ................................................................. 98
Table 19. Access and Egress Modes Frequency and Joint Distributions for 332,073 Trips ..... 102
List of Figures

Figure 1. Traditional Framework for Evaluating Policies and Strategies................................. 6
Figure 2. "Reverse Engineering" Framework for Evaluating Policies and Strategies................. 7
Figure 3. Map of the City of Toronto and the GTA................................................................. 9
Figure 4. General DTA algorithmic procedure (TRB, 2011) .................................................. 24
Figure 5. Constraint programming procedure (Beck, 2012) .................................................... 29
Figure 6. Travel Time Components.......................................................................................... 55
Figure 7. Network Representation............................................................................................ 60
Figure 8. Passenger Path Set.................................................................................................... 64
Figure 9. The SORTAH Procedure............................................................................................ 72
Figure 10. The SORTAH Prototype as an Add-On Feature of MILATRAS .............................. 79
Figure 11. Case 1 – Shorter Time Check.................................................................................... 81
Figure 12. Case 2 – Measure of Unfairness Check................................................................. 82
Figure 13. Case 3 – No Travel Time Check ............................................................................. 83
Figure 14. Comparing Crowding Levels for each Case......................................................... 88
Figure 15. Comparing Total Travel Time for each Case....................................................... 90
Figure 16. Comparing Execution Time Relative to the Network Size for each Case.............. 92
Figure 17. GTA and Surrounding Regions for Year 2001 (Joint Program in Transportation, 2003a) ......................................................................................................................... 98
Figure 18. Number of Overcrowded Routes Before and After Re-Routing Process............... 105
Figure 19. Passenger Load Factor Profile and LOS Standards of Route 102 (Southbound) for the Before and After Re-Routing Process from 7:00 am to 8:00 am................................................. 107
Figure 20. Passenger Load Factor Profile and LOS Standards of Route 102 (Northbound) for the Before and After Re-Routing Process from 7:00 am to 8:00 am

Figure 21. Final SORTAH Re-Routing Procedure
Chapter 1
Introduction

1 Chapter Overview

This introduction chapter (Chapter 1) provides the background and discussion of the problem statement investigated by this thesis. It introduces the scope, objectives and motivation of this study. The chapter is organized as follows:

- Section 2 describes the background and discusses the problem statement of this thesis.
- Section 3 presents the motivation of this research.
- Section 4 highlights the scope and objectives of this study.
- Section 5 provides the structure of this thesis.

2 Problem Statement

Following World War II, the population of North American cities increased dramatically and the associated need for better mobility grew. In turn, urban motorization and the increased reliance on the automobile as the primary mode of travel accelerated remarkably. Though the growth in auto dependency has provided unprecedented levels of mobility and liberty to motorists, it also had adverse effects. The use of single occupancy vehicles has raised concerns about resource consumption, traffic congestion, and emission. The associated deteriorating levels of network gridlock reduced the economic, social and environmental viabilities of urban communities (Garber & Hoel, 2002; Litman & Laube, 2002). A study initiated by Metrolinx concluded that the annual cost of congestion in 2006 in the Greater Toronto and Hamilton Area (GTHA), Ontario, was $3.3 billion to commuters and $2.7 billion in the form of Gross Domestic Product
(GDP) to the economy. By 2031, it is estimated that these numbers will balloon to $7.8 billion and $7.2 billion respectively (Metrolinx, 2008; HDR Corporation, 2008). In such context, there is a growing interest in promoting sustainable communities that can benefit from more balanced multimodal transportation systems and few market distortions that favour automobile travel (Litman, 2002).

Transportation planning is an important sub-field of urban planning that plays a role in shaping our quality of life. The decision to invest in transit, highway, or pedestrian infrastructure greatly affects the activity patterns of individuals and the movement of goods, because transportation systems are an integral component of an urban area’s social, economic, and physical structure. It typically consists of a set of networks serving various travel modes such as public transit and the private automobile. By comparison with the amount of research on policy and planning evaluation or management tools of automobiles on roads, there has only been a small amount of research related to public transit. Recently, public transit and transit planning has become an important field of study as concern for the environment and the sustainability of urban areas is rising. Public transit is becoming a competitive and attractive alternate mode of travel because it enables high capacity, energy efficient and low emission movement of people. It is widely considered a cost-effective alternative that mitigates the effects of traffic gridlock, providing auto owners who do not want to drive with a travel alternative and presenting an essential service for those who lack access to private vehicles (Hale, 2009). This is especially true for public transit systems that are implemented with emerging information, communication, and sensor technologies that enable innovative transit operations control strategies and customer information systems such as Advanced Public Transportation Systems (APTS) and Automated Travellers Information Systems (ATIS).

Transit assignment is one of the stages of transportation planning in predicting passenger loads and level of services (LOS) on a given transit network. Models based on transit assignment are widely used as an important planning tool at the strategic and operational levels. They can be used to evaluate important decisions concerning investment in public transport infrastructure or services. However, traditional transit assignment methods are plagued with many problems.
They are generally aggregate, and therefore more appropriate for regional planning than community or neighbourhood planning. With the increasing use of Intelligent Transportation Systems (ITS) technologies on transit networks, existing transit assignment models have serious limitations in modelling the effect of real-time information provision on passenger travel choice behaviour. Traditional mode choice models assume rational passenger behaviour, complete knowledge of the transportation system, and perfect information about all the available alternate routes and their choice consequences (Simon, 1957; Barros, 2010). These assumptions poorly characterize actual human behaviour because rationality of passengers is bounded by the information they can have, the cognitive capacity of their minds, and the terminable amount of time available to them to make decisions (Simon, 1957; Barros, 2010). Human behaviour can change as result of personal interaction or observation of the environment through a learning process (Levy, 2012). Recent research has been focusing on modelling this adaptive travel behaviour of people in response to a change in their environment using advanced methodologies, such as the microsimulation approach, that are already well established in traffic assignment procedures (Wahba, 2009). These state-of-the-art transit assignment modelling methods provide a new dimension for improving current practice in transit service planning. They answer the questions of what the impact ITS technologies have on the performance of public transit networks, how deployment of ITS technologies can be modelled, and how the effect of ITS deployments on the performance of public transit networks can be measured. Yet, they do not provide transit agencies with the network operational strategy that they should be striving for when deploying ITS technologies. ITS technologies have an impact on passenger travel choice behaviour. How can ITS deployment be managed to better mitigate overcrowding of transit networks and improve the efficiency of transit network operations?

3 Motivation

Similar to road networks, transit networks also have congestion problems. Congestion on road networks is defined differently from transit networks. Road networks are plagued with traffic gridlock issues, while transit networks are plagued with overcrowding problems. Road networks
are more concerned with roadway capacity and the flow of vehicle throughput, while transit networks are more concerned with transit vehicle capacity and the service provided to allow sufficient passenger throughput. Overcrowding on public transit causes discomfort and deteriorates the service of transit networks. It is a challenge for transit operators to manage overcrowding to efficiently run transit systems.

The Toronto Transit Commission (TTC) network is a transit system of surface routes (such as buses and streetcars) and underground subway lines that serves the City of Toronto. Since 2006, the average daily ridership for TTC has been increasing at an annual compounded rate of 2%. This is equivalent to the 18% increase in the annual ridership for a 7-year period from 445 million passengers in 2006 to 525 million passengers in 2013 (TTC, 2013). It is projected that TTC ridership will continue to increase at a similar rate to 536 million annual passengers by 2015 (TTC, 2011). Though TTC has increased its network supply from 149 to 158 surfaces within the 7-year period, it is still insufficient to serve the increase in ridership (TTC, 2013). Transit passengers are still experiencing congestion and crowding on the TTC network, especially during peak periods.

As mentioned above, the direct cost of congestion to commuters in the GTHA was estimated to be $3.3 billion in 2006. Approximately $1.4 billion was estimated to occur in the City of Toronto, in which $2 million was associated with the delay to transit users (City of Toronto, 2014). In 2011, 847 complaints were received with regards to surface delays, which increased by 29% from the previous year with 659 complaints (TTC, 2011). These congestion costs and complaints will not go away altogether, because congestion has become an integral part of normal life in almost all large and growing metropolitan areas in the world (Downs, 2004). It performs a vital and necessary positive function in modern society (Downs, 2004). Thus, congestion cannot be eliminated but its adverse effects can be reduced through various policy tools.

Recently, urban planners and politicians have been increasingly alarmed by the deteriorating levels of network gridlock and its impacts on the quality of life. Policy agendas are typically promoted to build more roads, to expand transit, and to encourage high-density settlements to
help reduce the adverse effects of congestion, in response to the intensification of congestion. In 2009, the Government of Ontario committed $8.4 billion to support new transit for the City of Toronto as a part of The Big Move, a 25-year plan to transform the regional transportation network across the GTHA (Metrolinx, 2008). The new plan includes three rapid transit lines (which include the Eglinton Crosstown LRT, Sheppard East LRT, and Finch West LRT) running above and below ground to connect Toronto with comfortable, convenient, reliable and fast service (Metrolinx, 2008). However, increasing supply by building more infrastructure may not necessarily improve network performance and reduce congestion.

As per Braess’s paradox, adding extra capacity to the network when the moving entities selfishly choose their route can in some cases reduce the overall performance (Braess, 2005). This system is known as user equilibrium or Nash equilibrium, where agents will continue to change their routes to improve their respective travel times until they have no more incentive to do so. Due to the non-monotonicity property of user equilibrium systems with respect to the network’s capacity, passengers continue to change paths despite the reduction in overall network efficiency.

In addition, infrastructure can take 5 to 20 years to build before it becomes fully operational. This is especially true for transit infrastructure. The Eglinton Crosstown LRT commenced in the spring of 2010 and is projected to start operating in the winter of 2017 (Metrolinx, 2008). Modifications and improvements are scheduled to occur until 2022 before it becomes fully operational (Metrolinx, 2008). During this period, the projected ridership in which the LRT is designed for can change with the growth of the region’s population, citizens’ incomes, and general prosperity. Policies that increase network capacity alone would not improve operational efficiency. Anti-congestion tactics need to be applied in conjunction with capacity policies to slow down the rate at which congestion is worsening and allowing for better utilization of existing and future supply.

Since the 1970’s, transport planners have directed their focus towards managing the increasing travel demand rather than boosting supply to reduce auto congestion. This is known as Travel Demand Management (TDM). The purpose of TDM is to reduce travel demand by encouraging mode change or to redistribute the demand in space or in time. Managing demand can be a cost-
effective alternative compared to increasing capacity. It has the potential to deliver better environmental outcomes, improved public health, stronger communities, and more prosperous and livable cities. Measures in support of this strategy may include information provision and incentives for encouraging car-pooling that redistribute demand or parking restrictions and fuel price increases that change the relative level of service (LOS) in favour of other modes than the auto-driver. The transit congestion problem is characterized by overcrowding of several transit routes in a network, causing extra waiting time and ultimately travel time. By expanding the concept of TDM to the transit congestion problem, such as information provision via ITS technologies to encourage redistribution of demand from the over-crowded routes to the less utilized routes, transit networks can be relieved of “pinch points” that are causing the overall congestion.

Figure 1. Traditional Framework for Evaluating Policies and Strategies

As mentioned above, existing transit assignment models can be used to predict passenger loads and levels of services in a given transit network. More advanced models can respond to the impacts of emerging ITS technologies and allow transit operators and planners to experiment with different TDM policies and strategies. However, they do not provide transit operators and planners with the network operational strategy that they should be striving for when deploying ITS technologies. The traditional framework for evaluating policies and strategies is comparable to the forward engineering process of product development moving from high-level abstraction to the final product. In this case, the policy would be the abstraction and the final product would
be the resulting network performance. As shown in Figure 1, a policy or strategy is first developed and then tested for its impact via a transit assignment model. The model outputs performance measures that reflect the goal of the policy and then is used to compare the before and after scenarios. Depending on the outcome of the after-scenario, transit operators and planners would determine whether the policy or strategy would improve the current performance of the network. With this framework, transit operators and planners cannot determine whether their efforts are well justified when developing the policies and strategies. The upper bound, beyond which the network performance can no longer improve, is not known. By understanding what the bound is, transit operators and planners can better concentrate their efforts on developing effective policies and strategies.

Figure 2. "Reverse Engineering" Framework for Evaluating Policies and Strategies

The motivation is to enhance the current state-of-the-art transit assignment model to determine the upper bound beyond which network performance can no longer improve and demand cannot be redistributed any further to relieve the congestion. This supports the proposed “reverse engineering” framework for evaluating policies and strategies whereby a network solution would drive the development of a policy or strategy (shown in Figure 2).
4 Scope and Objectives

This thesis attempts to explore the following research question:

*What is the upper bound at which transit overcrowding is minimized while accounting for individual passenger travel needs?*

In other words, what is the threshold at which demand cannot be redistributed any further to relieve congestion? There are several approaches proposed in the literature that can determine this bound. Microsimulation methods apply predefined rules to estimate the resulting effects on the network. They are popular in practice because they can be sophisticated and can capture fine detail of the real world system that is to be investigated. However, these simulation models can be computationally demanding due to their complicated nature and the huge size of the underlying transportation network (Axhausen, 2011). Another method is through analytical models. They are built on mathematical equations and inequalities. They only concentrate on certain elements considered to be important for a particular analysis and simplify representations of the real world (Ortuzar & Willumsen, 2001). These problems are typically non-deterministic polynomial-time hard (NP-hard) due to the range of interrelated requirements on their components that need to be mathematically represented in order for the problem to be solved. In such context, a heuristic approach is proposed in this study.

Heuristics are techniques designed for solving a problem more quickly when classic methods are too slow, or for finding an approximate solution when classic methods fail to find any exact solution. This method has been used in transportation problems such as freight models, aircraft routing and scheduling, and train timetabling (Pinedo, 2009). However, it has not been explored in the transit assignment context. The heuristic approach is proposed to balance the traditional analytical methods and advanced microsimulation models used to solve the transit assignment problem.

The objective of this dissertation is to develop a transit assignment heuristic that minimizes the overall congestion effects of a transit network while balancing individual travel needs. It also presents a large-scale real-world application of the proposed modelling system for the TTC.
public transportation system within the context of the City of Toronto as a case study, shown in Figure 3.

Figure 3. Map of the City of Toronto and the GTA

[Map of the City of Toronto and the GTA]

The proposed heuristic is called the System Optimal Re-routing Transit Assignment Heuristic – SORTAH. It solves for a system optimal solution that is widely studied for road networks, but less considered for public transit systems. Traditionally, transit assignment models use user equilibrium routing methods that assume a selfish philosophy where passengers choose a path that would give them the most benefit. The assumption may be a good representation of actual passenger behaviour when distributed on an existing network, but it does not lead to the best possible use of the network that minimizes overall congestion effects. System optimal routing methods assume an altruistic philosophy where passengers would choose a path that would benefit the entire network. This gives a solution where demand cannot be redistributed any further to relieve the “pinch points” that are causing the overall congestion.
SORTAH is a heuristic that can only be applied to agent-based models. Individual passengers are represented by agents with passenger path sets. The traditional path choice model is replaced by a path distribution model that directs passengers to their path for the benefit of the entire network. Conventionally, system optimal models are solved by minimizing the total travel time, which does not necessarily reduce crowding levels on the network. The goal of SORTAH is to solve for a network with minimum number of “pinch points” or overcrowded routes in the network while balancing individual travel needs. This thesis hypothesizes that by directly minimizing crowding levels, total travel time will indirectly be minimized when solving for the user-constrained system optimal solution. Three variations of the SORTAH prototype are created to test the hypothesis and presented in this thesis. SORTAH adopts the most effective algorithm for re-routing passengers, which minimizes crowding level and travel time factors when re-routing passengers. This study is the first step in achieving the “reverse engineering” framework for policy and strategy development to reduce transit congestion.

5 Thesis Structure

The thesis is organized in the following manner:

Chapter 1 introduces the motivation for developing a heuristic to determine the upper bound at which the overall congestion effects of a transit network is at its minimal and the main research objectives.

Chapter 2 provides a summary of existing and theoretical transit assignment models as a literature review. It presents a brief history of the developments in the field of transit assignment from early developments to the recent advancements.

Chapter 3 reviews transit performance measurements and transit characteristics that provided basis for the conceptual development of the proposed approach.

Chapter 4 presents the SORTAH methodology, including the requirements, procedure, and analysis framework.
Chapter 5 presents the prototype implementations of the heuristic.

Chapter 6 reports on the large-scale application of the heuristic, discussing the data requirements and modelling assumptions.

Chapter 7 concludes with a summary of key findings, an evaluation of the SORTAH method, and future directions to advance heuristic application to transit assignment and the “reverse engineering” framework for policy and strategy development.
Chapter 2
Literature Review

1 Chapter Overview

This chapter presents an overview of the developments in the field of transit assignment from early developments to recent advancements.

- Section 2 describes the early developments of transit assignment.
- Section 3 reviews various studies on user equilibrium transit assignment.
- Section 4 reviews the limited studies on system optimal transit assignment. It also discusses the concepts and theories of the widely studied system optimal traffic assignment for road networks and its application to the transit network.
- Section 5 describes various complex problem solving techniques that can be used to solve transit assignment problems.
- Section 6 provides an overview of the literature reviewed and the author’s vision for the future developments in the area of transit assignment modelling.

2 Early Developments

A significant portion of transit assignment research efforts were based on the user equilibrium (UE) paradigm. The concept of user equilibrium was first described by Wardrop in 1952 to model users’ route choice in auto traffic network. The model’s mathematical equations were formalized by Beckmann, McGuire, and Winsten (1955). The formulation allows you to find user equilibrium link flows by solving an optimization problem. Travellers are assumed to know the precise route travel time and choose the fastest route. When UE is reached, all the used routes
of an origin-destination (OD) pair would have the same and shortest travel time; no traveller can reduce his/her travel time solely by changing route and the equilibrium based traffic distribution would stay stable (Wardrop, 1952). This core assumption dictated many traffic assignment models. In other words, the UE in an auto traffic network is analogous to the typical market equilibrium in neoclassical economics, and many traffic assignment objectives have focused to find this equilibrium state. The UE concept was then extended to the transit network.

Prior to the UE concept, transit assignment typically followed the all-or-nothing assignment in which all the passengers of an OD pair would choose the shortest path. The shortest path was determined by shortest-path algorithms and later the strategy or hyperpath approach that became commonly adopted in many subsequent transit assignment models. These path searches were also adopted in UE-based transit assignment.

2.1 Shortest-Path Algorithm

In the shortest-path algorithms, the journey cost was minimized according to the path they use. Dial (1967), Fernside and Draper (1971), and LeClerq (1972) developed the early shortest-path algorithms. It was assumed that deterministic vehicle-running times and waiting times depended on the frequencies of routes serving the O-D pair. For example, Dial was one of the first ones to present the first transit assignment model. The network was modelled by “trunklines” that had elements of origin, destination, travel time and a set of lines serving the arc. He explains that the standard shortest path algorithm by Moore (1957) was adapted for transit because the Markov property is violated. The Markov property states that the route choice probabilities at a node are independent of the travellers’ origin, which is not true because there are only a certain number of routes that are available at each origin stop. An issue with Dial’s representation of the network is that it does not deal with situations where lines connect the same OD pair but have different travel times. To address this, Fearnside and Drapper (1971) represented each node by line specific nodes and boarding and alighting links. This allowed for the inclusion of waiting times and line specific transfer penalties that were not reflected in the earlier network representation of Dial’s work. However, this created multiplicity of lines and nodes in the network representation. It also lead to an overestimation of waiting time if it is realistic to assume that passengers choose
the first vehicle arriving from several possible common lines. Since it is assumed that passengers have to decide which node for a specific line to board, it creates the common line problem that is unique to transit assignment. To solve this problem, Le Clerq (1972) developed an algorithm that searches for the shortest by looking at all possible interchanges from the service the traveller is currently using.

The use of these shortest-path heuristics that were originally developed for user equilibrium-based traffic assignment problems was soon discovered to be insufficient in the context of a transit network. Although there are similarities between auto and transit networks, there are still many differences between the two types of network making transit networks not directly comparable to auto networks. First, link travel time is typically the only variable affected by congestion in auto networks. But in transit networks, overcrowding have other negative externalities upon route cost such as in-vehicle discomfort, extended vehicle dwell time, and failure to board due to vehicle capacity constraints which increases waiting time. Transit networks have an additional problem known as the common lines problem, which poses more complexity in overcrowding impact. The common lines problem occurs when multiple lines connect the same OD pair but have different travel times (e.g. express service versus regular service). When a passenger is at a transit stop that is shared by several competitive transit lines, they have a choice between boarding a vehicle from one line or waiting for a vehicle from another express line to minimize their individual total travel time. It was difficult to determine which vehicle the passenger would choose to board. To address this issue, the notion of strategy and hyperpath was introduced (Spiess, 1984; Nguyen and Pallottino, 1988).

2.2 Strategy and Hyperpath

The notion of strategy was first introduced by Spiess (1984). It was later formalized and formulated as a linear programming problem in a space of line flows by Spiess and Florian (1989). It was a ground breaking work because it established a fundamental assumption for transit users’ route-choice behaviour in many of the later transit assignment models. It states that passengers developed strategies if more than one transit lines lead to their destination from their current location. ‘Strategy’ is a choice of an attractive set of lines at each boarding-decision
point; it was created by adopting the theory of people’s choice behaviour by Lampkins and Sallman (1967), in which passengers would exclude lines from the choice set if they are obviously bad. Passengers would board the first arriving vehicle from the set of attractive lines, significantly reducing their travel time. By formulating the problem as a linear program in a space of line flows, the probability of choosing different lines was determined. It was shown that employing the optimal strategy would minimize passenger’s travel time, satisfying the user equilibrium requirement.

Similar to the notion of strategy, a graph-theoretic language known as hyperpaths was introduced by Nguyen and Pallotino (1988). A hyperpath is a set of paths that could potentially be taken by the traveller depending on the arrival of the first vehicle at the passenger’s origin and the station in which he or she has to interchange. This problem is reduced to finding the shortest hyperpath, or the set of paths that should be included in the traveller’s route choice, because it assumes that the probability of taking a path is given according to line frequencies. It also assumes the underlying passenger behaviour to be such that before a passenger starts his/her trip, ‘a passenger has chosen a fixed subset of transit lines for every stop he may encounter on his trip and for every transit line the alighting point’ (Nguyen and Pallottino, 1988). The solution of the hyperpath model is the same as the Spiess and Florian’s strategy-based model. However, an additional advantage of the hyperpath model is that it has more flexible network description that allows for more efficient computational techniques on larger networks. For example, the hyperpath model helped extend Dijkstra’s shortest-path algorithm into a “shortest hyperpath algorithm” (Comminetti and Corre, 2001). The waiting time function of flow was formulated as a simplified bulk queue model of Gendreau (1984) and the UE transit assignment was in the space of line flows, similar to the hyperpath model.
3  User Equilibrium Paradigm

Wardrop’s first principle is such that “journey times on all the routes actually used are equal and less than those which would be experienced by a single vehicle on any unused route.” Traffic patterns that satisfy this principle are called user equilibrium (Dafermos and Sparrow 1969).

The UE-based transit assignment model was one of the first models to represent users’ route choice. The static UE model is the classical transit assignment model. It assumes demand and supply are constant within the specific period of analysis and therefore commonly employs frequency-based network formulations. But in reality, there are time-dependent fluctuations of demand that results in congestion effects due to vehicle capacity constraints of each transit line. To accommodate all issues including crowding, capacity problems, and passenger’s choice behaviour, the static UE model was extended to stochastic user equilibrium models and dynamic user equilibrium models. This section discusses the development of the user equilibrium concept and describes the three main types of UE models: classical (static), stochastic, and dynamic.

3.1  Classical Static User Equilibrium

In the early research efforts of transit assignment, the classical UE approach shaped many transit assignment models because of the similarities identified between auto networks and transit networks. For example, travellers would always choose the most cost-efficient path to complete their journey in both networks. The classical UE model was an abstraction of the real world with three simplifying assumptions. First, travellers are assumed to choose the available route having the least travel time between their origin and destination. This reflects the idea that travel is rarely a goal in and of itself, but involves some time, cost, or disutility in which travellers would prefer to minimize. Since travel time of a route between an OD pair also depends on the choices made by other travellers who are also trying to choose the least travel time route between their own OD pair, the UE condition is achieved when every traveller succeeds in finding a route such that every used route has the minimum time or cost between the OD pair. Second, it assumes that travellers know and accurately perceive travel times through the network. This perception is presumably gained through numerous trials of different routes to be able to identify their
minimum travel time route given any congestion scenario. Third, the OD flows and network supply characteristics are assumed to be fixed and known. A solution that satisfies the three underlying assumption will result in a static UE-based transit assignment. (TRB, 2011)

With the combination of the three assumptions, notably the latter two, the classical UE transit assignment model is inherently static in nature. Passengers do not necessarily have an accurate perception or full knowledge of the transit network. Despite these unrealistic assumptions, the UE concept is still a dominant paradigm in transit network analysis. Firstly, the assumption about passenger’s selfish route choice behaviour is reasonable and can lead to efficient solution methods that yield transferable conclusions. Other simpler behavioural assumptions such as all passengers choosing the same shortest route do not lead to the same methods of quality. Secondly, since the modelling time horizon is often long enough to assume that most travellers have discovered the shortest routes for their trips coupled with the recent advances in advanced traveller information systems (ATIS) and other ITS technologies, travellers can become much more aware of the network conditions and have a relatively accurate perception of the travel times than before. Finally, the UE model has strong mathematical foundation; methods from economics can be adopted for comparative analysis of the UE-based problem. These economic methods can evaluate the potential benefits that accrue to travellers following a change in travel conditions due to implementation of certain transportation projects or policies.

As mentioned, the static representation has notable advantages, especially allowing for mathematical properties of transit assignment models to be obtained relatively easily. The existence and uniqueness of an equilibrium state needs to be proven and a solution algorithm needs to be defined in order for a UE solution to be found. In the early UE-based transit assignment studies, they were proven and defined. For example, Wu et al. (1994) gave sufficient conditions for equilibrium’s existence and uniqueness, and a solution algorithm. The objective function needed to be strictly convex in order for uniqueness condition to be met. It was shown by Bouzaïne-Ayari et al. (1995) that however, the existence and uniqueness properties of an equilibrium typically required unrealistic assumptions about waiting time and travel time functions of flow. Later, it was proven by Cominetti and Correa (2001) that when weak
conditions are available, in which travel time functions of flow are not strongly monotonic, equilibrium properties can be obtained without giving a solution algorithm. Extending upon this work, Cepeda et al. (2006) newly characterized the equilibrium to formulate an equivalent optimization problem in terms of a computable gap function that vanishes at equilibrium. This allows for the generalization of strategy-based transit network equilibrium model and can address flow-dependent travel times.

Static assignment is ill-suited to analyze either traffic congestion effects at a fine-grained temporal level, or many of the measures that can be taken to address congestion. The following two sections discuss later research efforts that attempt to have a better representation of the real-world transit networks by relaxing the last two assumptions of the classical UE model.

3.2 Stochastic User Equilibrium

In the classical UE framework, travellers are assumed to have a precise perception of route travel time. However, this is an unrealistic assumption because passengers typically do not have full information of all the possible routes from the origin to their destination. To overcome this assumption, the original UE was extended to stochastic user equilibrium (SUE) by Daganzo and Sheffi in 1977. SUE applies randomness (a stochastic nature) into the user’s perceived cost when choosing routes by incorporating random utility maximization (RUM) choice modelling. Its solution is still obtained when no traveller’s perceived cost can be reduced solely by changing route.

The RUM choice model is a discrete choice analysis tool with basis in microeconomics theory and cognitive psychology. It uses utility that integrates attributes of choice alternatives and decision-maker social-demographic characteristics to determine one’s choice. The model was first introduced into travel demand analysis in the 1970s and became the backbone of many travel behaviour studies (McFadden, 2005). Most of these RUM choice models were originally applied in the context of analysing travellers’ mode choice rather than transit users’ route choice, focusing on the value of travel time savings, elasticity and willingness to pay in its utility functions. However, these parameters were irrelevant to route choices. In order to extend these
choice models into the transit users’ route choice context, the impacts of level-of-service (LOS) factors on passengers’ choice was emphasized. It assumes that decision-makers have a perfect discrimination capability of choosing the alternative with the maximum utility perceived; analysts typically do not have complete information about this utility and need to represent the uncertainty as a random term. Nielsen (2000) summarized four uncertain factors in which random terms can encompass for transit users’ route choice:

1. Passengers do not have full knowledge of the transit network and they only choose according to their perceived utilities,

2. Travel times along different routes may vary from day to day,

3. Different routes are often chosen for the sake of variation, and

4. Different passengers may have different preferences.

These factors can be incorporated into the utility functions for discrete route choice analysis of transit users. Examples of different RUM choice models include the logit model, probit model, and multi-leveled Nested Logit model.

Hunt (1990) was the first to propose a logit choice model (which is a type of RUM choice model) for transit route choice. The logit model was often extended and adopted in later SUE transit assignment models. For example, Lam et al. (1999b) presented a mathematical programming that was based on a multinomial logit route-choice model. The study proved that when the link-capacity constraints are reached, the Lagrangian multipliers of the mathematical programming were equivalent to the equilibrium passenger-overload delays in overcrowded transit network. This model was later extended by Lam and Xie (2002) to incorporate a dwell-time model from Lam et al. (1999a). On the other hand, Nielsen (2000) based his SUE transit assignment model on a multinomial probit route-choice model. Nielsen and Frederiksen (2006) proposed an SUE transit assignment using the Nested Logit route-choice model, while Yang and Lam (2006) put forward a probit-type reliability-based SUE transit assignment model. In Yang and Lam’s (2006) model, in-vehicle travel times were normally distributed. The disutility
function was used for considering passengers’ risk-averse behaviour under unreliable transit conditions, which is associated with passenger travel time and its variation.

RUM choice models are grounded in well-established microeconomic theories, but also with strong background in cognitive psychology as well. Given people are influenced by context effects, emotion, and errors in perception and judgment (McFadden, 2003), utility assessment of people’s choice is difficult. The SUE models have a strong basis in RUM models. Hazelton (1998) argues that for more plausible SUE, RUM may not be a good tool because SUE is a deterministic model by nature. It relates to a specific isolated state, which is guaranteed by the Law of Large Number. SUE only considers the probabilities of passenger choice, but does not model how networks react to these choices. The following section addresses both issues of demand and supply dynamics.

### 3.3 Dynamic User Equilibrium

In the static assignment framework, demand and supply are assumed to be constant within the specified period of analysis. But in reality, traffic and transit networks are generally not in steady state. During the peak-period, transit systems usually experience significant loading where capacity is exceeded, affecting passenger satisfaction (Lam et al., 1999a). Time-dependant demand and departure time choice strongly contribute to dynamics in passengers’ route choice and network assignment. As a result, bottleneck-induced overcrowding problem cannot be revealed.

In order to represent time variations in flows and network conditions, research efforts started to focus on dynamic transit assignment in the late 1970s. To retain the advantages of an equilibrium approach while accommodating the dynamics of route choice and network model, the notion of UE needs to be extended in two ways. First, the static model’s perfect traveller information assumption and route choice criterion needed to be generalized. This generalization recognizes that travel times on network links vary over time. In the static framework, link travel times will prevail at the instant of departure from the origin; in the dynamic framework, experienced link travel times depends on when they arrive at various links along a route and on the travel times
that prevail on the links at those specific future times. Second, the UE condition of equal travel
times on used routes only applies to travellers who are assumed to depart at the same time
between the same OD pair. This recognizes that travellers who depart from an origin to a
destination at different times will experience different travel times. These extensions mark the
fundamental departure from static assignment in network representation and algorithmic design.

Many dynamic transit assignment studies have focused on finding an assignment in the network
that has a given known time-dependent OD trip pattern. For example, Tong and Wong (1999)
presented a dynamic stochastic transit assignment that used a schedule-based model with given
time-dependent OD matrix. A branch-and-bound algorithm that was based on the work of Tong
and Richardson (1984) was used to generate the time-dependent optimal path and a Monte Carlo
approach was employed to solve the stochastic assignment problem. Poon et al. (2004) also used
a schedule-based network with given time-dependent OD demand for their proposed dynamic
UE transit assignment. It assumes that transit vehicles with capacity constraint operate precisely
as scheduled and passengers queue according to the single channel first-in-first-out (FIFO) rule.
Each vehicle’s available capacity is dynamically updated as demand is loaded onto network
through time-increment simulation. The same branch-and-bound algorithm was used to update
the optimal paths and used in the next simulation run. The method of successive averages (MSA)
was used to solve the assignment problem. Both studies conducted a case study based on data
from the Mass Transit railway System in Hong Kong.

In a study by Hamdouch and Lawphongpanich (2008), a schedule-based UE transit assignment
model with vehicle capacity constraint was presented. In this model, passengers’ travel strategies
were represented by sub-graphs and they were adaptive over time. Vehicle loading followed
rules such that on-board passengers have priority, waiting passengers are loaded by a FIFO rule
or in a random manner, and passengers unable to board must wait for the next vehicle. The UE
flow model was formulated as a variational inequality involving a vector-valued function of
expected strategy costs. The assignment problem was solved using MSA and a dynamic program
that took successive averages and generated strategies for each iteration; the algorithm converged
to an equilibrium solution.
As mentioned, most dynamic transit assignment studies commonly use schedule-based networks. Schmocker et al. (2008) was the first to introduce a frequency-based dynamic transit assignment model for an overcrowded high-frequency transit network. Passengers that cannot board a vehicle would remain on the platform. Passenger’s attitude of risk averse was reflected by the introduction of a fail-to-board probability. This model adopts Nguyen and Pallotitino’s (1988) hyperpath search algorithm; the expected delay along a hyperpath was represented by the fail-to-board probability and was incorporated into a generalized cost. The route-choice model partially considers the dynamic effects where it is assumed that passengers expected current level of crowding remains constant in future time intervals. The chance of failing to board in future time intervals remains the same as in the current interval. The assignment problem was also solved using MSA and a case study was conducted for the London underground.

The equilibrium concept is generally applied to route choice, but it can also be extended to passengers’ departure time choice. For a network with single origin and single destination, Huang et al. (2004) proposed an equilibrium formulation of bus riding in the morning peak. An overcrowding function that made trade-offs between overcrowding cost and early/late arrival penalty at destinations to determine passengers’ optimal departure times was developed. On the other hand, for a many-to-one mass transit system, Tian et al. (2007) further analyzed the equilibrium properties of the morning peak. A trade-off between the combination of travel time and overcrowding cost against the schedule-delay cost was used to determine the optimal departure times from various origins to a single destination. At times, departure time choice and route choice can be of equal importance to the passenger. Hamdouch et al. (2011) proposed a schedule-based equilibrium model that differentiates the discomfort level experienced by sitting and standing passengers. The problem is formulated as a variational inequality that involves a vector-valued function of strategy costs. At each iteration of the dynamic program, MSA is used to generate strategies for solution finding.

When timetables are reasonably reliable and service frequencies are low, the equilibrium concept can also be applied to a model where departure time and route choice are simultaneously considered. Nguyen et al. (2001) proposed a transit assignment that follows the concept of path
available capacity to encompass departure time and route choice at the same time. It captures the flow priority induced by the first-in-first-out (FIFO) rule that binds passengers at every access point to transit network. Capacity constraint was imposed by a boarding penalty cost function, which is dependent on the difference between the queue length before the boarding passengers and the capacity of the vehicle being boarded. The UE conditions were mathematically formulated. The assignment problem was solved using an algorithm developed based on the boarding penalty cost functions.

Overall, dynamic use equilibrium transit assignment models are similar to dynamic traffic assignment models. They generally follow a sequence of three algorithmic components iteratively until a defined stopping criterion is met to find an equilibrium solution (TRB, 2011). The first component is the route evaluation step. The effects of passengers choosing specific routes are determined through a network loading process. This can be done through an analytical or simulation-based network loading approach. Analytical models use flow functions to predict how passengers propagate in the network; simulation-based models use a mesoscopic simulation approach to represent changes in the network flow at resolution of 5 to 10 seconds. The second component is the path set update. The result of the network loading is analyzed. The route with the lowest experienced travel time between every OD pair, for each departure time period, or assignment interval, is found by a time-dependent shortest path (TDSP) algorithm. The path set is updated by adding the newly found TDSP to all TDSPs found in previous iterations for the same OD pair and departure time. The third component is path assignment adjustment. The purpose of the adjustment is to achieve equal travel time among all routes in the current set. Passengers can shift their route choices towards the least experienced travel time route and the assignment can be brought closer to equilibrium. Once path assignment is adjusted, the algorithm returns to the first component, route evaluation, to determine the resulting transit network conditions. The stopping criterion is typically computed at the end of the network loading step (component one). Mathematical approaches such as MSA can be employed for the stopping criterion to impose convergence. The general dynamic transit assignment procedure can be found in Figure 4.
4 System Optimal Paradigm

According to Wardrop’s definition of the second principle of equilibrium, system optimum occurs when drivers cooperate with one another in order to minimize total system travel time. In effect, it is a measure of how inefficient the user equilibrium solution is by measuring the flow’s total travel time. When congestion effects are ignored, where link travel times are no longer a function of link flow, UE and SO solution methods produce identical results (Sheffi, 1985). This is because the UE objective function becomes the SO objective function, which is the minimization of total travel time. In effect, this becomes the all-or-nothing assignment because congestion would not be present and all flow would be assigned to the single path with minimum travel time; the resulting flow pattern would be both UE and SO solutions. The reason for all-or-nothing assignment to be invalid is that travel time increases on a path as the number of people taking the path increases; this endogenously models congestion.

In general, UE methods assign passengers to routes for each OD pair such that they experience the same travel time; no unused routes would have a lower travel time. On the other hand, SO
methods would assign passengers such that their paths have equal marginal travel times; no unused routes would have a lower marginal travel time. Since route travel times experienced by passengers for any given OD pair is different, the mean travel time is flow-weighted in SO solutions. It has been found that when flow over the network is relatively low, UE and SO flow patterns are similar (Sheffi, 1985). This is because network links are not congested and travel times are almost insensitive to additional flows. As flow along the paths increase, the UE and SO flow patterns become increasingly dissimilar (Sheffi, 1985). When the amount of flow is near the path’s capacity, then congestion is present. Typically, if SO problems are solved by marginal travel times, UE problems are solved by actual travel times. When congestion occurs, the marginal travel time between paths become very large even though the travel time itself is bound by the capacity; solving the same problem with marginal travel times rather than with the actual travel times become very different as congestion increases. For example, assume that there is a network of two routes in which one is a freeway and the other is a city street. If the total OD flow, $q$ is assigned to the freeway and the resulting travel time on the freeway is still less than the city street, then this would be the UE solution. However, if the derivative of the freeway’s travel time is larger than the city street’s derivative at this assignment, then the solution is not optimal from the system perspective. In the SO solution, flow needs to be reassigned to the city street such that even if the travel time of one driver increases, the travel time experienced by each remaining driver on the freeway will decrease.

It is important to understand that SO solutions produce normative flow patterns while UE solutions produce descriptive patterns. When extra capacity is added to the network to relieve congestion in an attempt to achieve system optimum, but flow is still distributed according to UE objective function, the Braess’s paradox can result. In other words, adding an extra route to the road network with the intent to reduce the SO objective function while distributing the flow according to UE objective function would not necessarily decrease the total travel time and each passenger’s travel time on each route. The theory behind the policy implementation needs to be consistent with flow distribution assumptions. Restricting travel choices and reductions in capacity may lead to overall flow distribution patterns; an example would be traffic control schemes such as ramp metering on freeway entrances.
There has been limited research in transit assignment under the system optimal paradigm. This is because most studies concentrate on modelling passenger choice behaviour on transit networks (Liu et al. 2010). Since passengers are the main users of transit, it is paramount to model their behaviour. However, it is often difficult to model choice behaviour given human inconsistency and variability. Transit assignment models are typically user-focused in order to understand the passenger behaviour response to transit networks and forecast future demand and network congestion manifestation. However, it is also important to use these models to evaluate potential improvements of transit networks that are typically congested due to capacity constraints. Congestion causes inefficiency in transit networks, which in turn will affect passenger behaviour on the network. Understanding the system optimal solution that attempts to minimize total travel time can improve efficiency and reduce congestion on the network.

Ma and Lebacque (2013) were the first to develop a dynamic system optimal routing model for a multimodal transit network. The model adopts the problem formulation of Peeta and Mahmassani (1995) to represent the path-based system optimal dynamic user equilibrium transit assignment problem. Passengers are assumed to follow the FIFO principle when boarding. The generalized travel cost consists of access time from the origin to the origin stop, transfer time between stops, boarding and alighting time, in-vehicle travel time, waiting time at stops, and egressing time from destination station to the destination stop. A cross entropy method is employed to solve transit assignment, compared to the traditional MSA method. The objective is to minimize the total experienced travel time. This is done by taking the time-dependent path marginal cost that represents the increase of experienced generalized travel cost of the system, as the incremental step in the solving algorithm.

There is a research gap in developing different methods to solve the system optimal transit assignment. In effect, the system optimal problem assumes that all passengers are controlled by the transit operating agency. In other words, the transit operating agency is scheduling each transit user to minimize the total travel time across the entire system. Methods used in scheduling could be investigated as a new approach to solving the system optimal transit assignment.
5 Complex Problem Solving Techniques

5.1 Mathematical Programming

Given the solid mathematical foundations of UE models, these problems can often be modelled as an optimization problem. When Beckmann, McGuire, and Winsten (1952) formalized the problem into mathematical equations, it was possible to find user equilibrium link flows by solving it as an optimization problem. Basic optimization techniques could be applied, but they would not be efficient as transportation networks grow larger. Large-scale equilibrium problems require an iterative approach, starting from an initial assignment then moving closer to the equilibrium solution by repeating a certain set of steps until it is ‘close enough’ to the desired solution. In general, equilibrium solutions repeat the following three steps:

1. Find the fastest path between each origin and each destination.

2. Shift travellers from slower paths to faster ones.

3. Recalculate link volumes and travel times after the shift, and return to step one unless the solution is close enough to the desired equilibrium solution.

The different commonly used mathematical models in transit assignment are: method of successive averages (MSA), Frank-Wolf (FW) method, and gradient projection (GP).

5.1.1 Method of Successive Averages (MSA)

The MSA is the simplest equilibrium solution algorithm. According to the procedure framework outlined above, the first and third steps are the same for all the equilibrium solution algorithms. MSA helps determine how travellers shift to the shortest paths found by finding a steady state solution. Initially, a lot of passengers are shifted. As the algorithm progresses, fewer and fewer passengers are shifted until we settle down on the average. MSA uses fixed step size as the algorithm progresses, which is one of the biggest drawbacks of the method. The purpose of MSA is so that both problems of shifting too few and too many are avoided. This method requires very little computer memory, but convergence to equilibrium solution is typically very slow.
5.1.2 Frank-Wolfe (FW)

The MSA method and FW method are similar. Rather than a fixed step size, FW algorithm uses an adaptive step size. At each iteration, the right amount of flow to shift in order to get as close to equilibrium as possible is calculated. The FW method also requires very little computer memory and is faster than MSA, but convergence is still slow for large networks. It has been the traditional algorithm for solving equilibrium problems. But it is shown to be inefficient for larger networks, which is usually the case for a transit network. More sophisticated algorithms with more memory and faster convergence is needed.

5.1.3 Gradient Projection (GP)

The GP method is basically a “path-based” algorithm, which was developed by Chen et al. (1998). It tries explicitly to track which paths each OD pair uses, and how many travellers choose each path. Since it is computationally unreasonable to track every possible path that could be used, the algorithms only track the paths which an OD pair actually uses. The algorithm is such that a set of routes is initialized for all OD pairs. The shortest path for each OD pair is found and added to the set if it is not already used. Within each OD pair, travellers are shifted to get closer to equilibrium. Travel times are then updated and paths that are not used are removed from the set and steps are repeated starting from finding the shortest path for each OD pair. In this algorithm, it is important to note that each OD pair is treated independently. In MSA and FW methods, the same step size would be applied across all links. If one OD pair is closer to equilibrium compared to another pair, there would be a finer adjustment to the first OD pair and larger adjustment to the latter. But in the link-based method, there is typically no such fitness and the same step size is bluntly applied to all origins. In other words, the step size would become very small after a few iterations because the method cannot move far from the equilibrium without disturbing an OD pair which is already close to equilibrium. However, many path-based algorithms differ in this aspect. The GP method uses Newton’s method to estimate the right amount of flow to shift from each path, which is longer than the shortest path to the shortest path. Overall, it can find more accurate solutions than MSA and FW methods in a faster time, but requires a large amount of computer memory.
5.2 Constraint Programming

Constraint programming (CP) has been a growing method of solving many industrial problems. Early constraint programming languages offered three essential components of CP: declarative problem modelling, propagation of the effects of decisions, and efficient search for feasible solutions. The search for solutions has strong basis in artificial intelligence (AI). Different search methods include generate and test (Golomb & Baumert, 1965), branch and bound (Lawler & Wood, 1966), the A* algorithm (Hart et al., 1968), iterative deepening (Korf, 1985), and tree search guided by the global problem structure (Freuder, 1994) or by information elicited during search (Maruyama et al., 1992), and by intelligent backtracking (Kondrak & van Beek, 1997).

There are three main features of CP: modelling, propagation, and search. Modelling involves the constraint representation of the problem to be solved at hand. It presents the problem in such a way that constraints must be satisfied for feasible solutions to be available. Propagation is a form of immediate feedback for decision-making which allows a system to spontaneously produce the consequences of a decision. Search is the important area for solving hard combinatorial problems which goes through a tree of solutions to find a feasible one. Figure 5 below depicts the general constraint programming process.

Figure 5. Constraint programming procedure (Beck, 2012)
The main focus of CP is placed on constraints and feasibility. First, feasibility information needs to be presented by defining elements with a domain store. If a variable has only one possible value, then it is pruned or propagated to only take that value. For variables that have more than one value, constraints are tested to determine if it is feasible with respect to the domain store. Given a solution space for the problem, pruning or propagation occurs when one carefully examines constraints of the program model to reduce possible variable values, eliminating infeasible configurations. Once obvious components of a solution are removed, branching occurs by decomposing the problem into sub-problems using heuristics or a search algorithm. Based on feasibility information, the search algorithm keeps adding constraints to domain stores one by one and solves. The search process needs to specify the tree to search and how to explore that tree. If the search does not find a feasible solution in the tree, it goes back to propagation and repeats the same procedure until a feasible solution can or cannot be found. It is important to note that constraint programming is mostly concerned with feasibility of a solution. It does not necessarily lead to an optimal solution. In order to find optimal solutions, it adds a constraint such that future solutions must be better than preceding solutions. The process is repeated until infeasible solutions are found so that the last solution found would be optimal.

CP has been applied to many applications such as planning, scheduling, resource allocation, and routing. Sports scheduling is a complex problem that involves many constraints and rules that games must be organized. For example, double round robin is used in sports scheduling problems. Teams are not allowed to play more than twice at home or away. Teams also cannot play another team more than twice. This problem is commonly solved in three steps. First, feasible strings of home away pattern sets are found. Then, games are assigned by matching the home away games accordingly. Lastly, times are assigned to the patterns. Following the three-step algorithm in a real world example, if there were 18 games per team, then there would be 38 different home away patterns. There would be 17 feasible pattern sets as a result. When games are assigned, 826 timetables would be generated. But when teams are actually assigned to the patterns, there would be 17 feasible schedules. To choose a final schedule, different choice criteria can be selected.
Travel costs contribute to a substantial part of the overall costs of sports tournaments. To minimize this cost, methods have been developed to find schedules that minimize total travel distance. Easton et al. (2001) introduces the travelling tournament problem that captures difficulties when scheduling tournaments with breaks and distance aspects. Given a number of teams, a distance matrix specifying the distance between the venues and an upper bound on the number of consecutive home and consecutive away games, a feasible schedule without repeaters that minimizes the total distance travelled by all teams is found. Many solution methods have been developed to solve this problem. Typically, a decomposition approach is used to separate the problem into two subproblems. Trick (2001) suggests that the most constrained subproblem should be solved first. This allows for more attention to the problem and effective solutions methods can be obtained.

Alternatively, Rasmussen and Trick (2006) proposes a timetable constrained distance minimization problem. A sports schedule is created by minimizing total travel distance by all teams. The timetable for a double round robin tournament with $n$ teams and represented by a distance matrix specifying distances between venues and an upper bound on the number of consecutive home and consecutive away games. Given a timetable for a double round robin tournament with $n$ teams, a feasible solution that minimizes the total distance travelled by all teams is found. A hybrid CP and integer programming (IP) method was proposed to solve the problem, and compared to CP and IP methods by themselves. In the hybrid method, all feasible patterns were generated using CP given a timetable. Then, the optimal pattern set by assigning each team to a pattern from the first step is determined by IP.

The decomposition of the sports scheduling problem can potentially be applied to transit assignment. Given the supply and demand interactions of transit assignment, decomposing the problem into supply and demand may lead to efficient solving of the problems. However, given the transit network size, constraint programming may be difficult.
5.3 Heuristics

Heuristics are experience-based techniques for problem solving, learning, and discovery that find solutions that are not guaranteed to be optimal. They speed up the process of finding a satisfactory solution via mental shortcuts to ease the cognitive load of making a decision where exhaustive search is impractical. When applying heuristics to solve a problem, it is important to understand the objective and purpose of situation and formulate the problem accordingly (Michalewicz & Fogel, 2004).

Heuristics are commonly used in scheduling, which is the allocation of resources to objects being placed in space and time. The process is subject to constraints and resources are allocated in such a way as to satisfy a set of desirable objectives. Specifically, educational timetabling is a common application. There are five sub-problems in the course timetabling problem: teacher assignment, class-teacher timetabling, course scheduling, student scheduling, and classroom assignment. Gunawan et al. (2007) proposes a heuristic to solve a university course timetabling problem that addresses the teaching assignment and course scheduling problem. The heuristic comprises of three main phases: pre-processing, construction, and improvement. During the pre-processing phase, course preferences and time periods preferences are assigned to each teacher. Then in the construction phase, a feasible solution is constructed to accommodate as many course and time period preferences as possible. From this initial solution, two operations are performed to seek further improvement of the solution during the improvement phase. Teachers are first re-allocated to courses and course sections and then re-scheduled to other days and time periods. Improvements can only be implemented if there is no violation of the maximum load constraint. The next movement is selected by exploring the possible neighbourhood in this phase. (Gunawan et al., 2007)

Another heuristic application of the timetabling problem is the proposed case-based heuristic selection approach for automated university course and exam timetabling (Burke et al., 2006). The goal of this research was to develop a timetabling system that was fundamentally more general than the current state of the art. It used knowledge discovery techniques to solve the problem. The problem and problem solving situations were modelled with specific heuristics of
the sub-problems. Then, the base case is refined and the cases that are proven to be non-useful to solving the new problem are discarded.

These timetabling problems are NP-hard. The allocation of resources to objects with space and time is complex with constraints of preferences and avoidance of conflicts. User-constrained system optimal transit assignment is essentially an allocation problem with various constraints. The development of a heuristic using existing research knowledge and experience can assist in finding a satisfactory solution without the requirement of powerful computational power.

6 Discussion

The literature review presented shows that there has been little effort in public transit research to model system optimal-based transit assignment. User equilibrium-based models have received the primary focus in transit assignment research because of their user-focused nature. These models were typically descriptive rather than normative to allow for validated demand forecasting on the network. However, it is important that normative demand modelling is applied to understand the extent to which network efficiency can be improved. By presenting a SO solution, researchers can potentially develop algorithms that move the current UE-based solutions to a more efficient and less congested transit network. The underlying assumptions between UE-based and SO-based models will need to be different because as pointed out from Braess’s paradox, UE concepts for flow distributions may not necessarily be applicable to a SO problem. It was shown that applying SO policies on a traffic network while applying UE flow concepts increased the total travel time of the system, decreasing the overall efficiency. To prevent this issue when solving for SO solutions, the supply should be kept constant, while representing how supply would react to demand changes in reality; in other words, congestion and capacity effects need to be modelled. It would not be possible to adopt UE assumptions for SO problems and the definition of attractive set would need to be redefined. Although UE solutions methods have been applied to the SO models of roadway networks, it may not be applicable given the nonlinearity of link flow functions of transit assignment. New solutions
methods from other applications such as scheduling can be adapted to transit assignment. Given the success that heuristics has had in solving complex timetabling problems, its extension into transit assignment is explored here.
Chapter 3
Transit Performance Measurement System

1 Chapter Overview

This chapter presents a performance measurement system developed as a part of the transit assignment heuristic to re-route passengers and to gauge how effective the network solution is at minimizing overall transit congestion.

- Section 2 discusses the purpose and importance of performance measurement systems.
- Section 3 describes the relevant stakeholders who are interested in transit performance.
- Section 4 reviews the different types of performance measures.
- Section 5 outlines the components of the SORTAH performance measurement system that serves as a basis for the proposed heuristic.

2 Purpose of Performance Measurement Systems

Transit performance has a substantial impact on people’s daily lives and upon the cost of providing transit service. Measuring the performance of a transit system is the first step towards efficient and proactive management. Transit agencies are required by regulatory bodies to report a number of performance measures. In the United States, transit agencies are required to report a number of performance measures to the National Transit Database as condition of receiving federal funds. These measures help decision-making bodies, such as transit boards and funding bodies, to make decisions on where and when services should be provided and to support actions designed to improve performance. However, performance measures are not collected and reported solely to satisfy regulations. Performance measurement data are also used for self-
improvement and communicating results of a transit agency’s management of the transit network to its constituents.

Transit agencies typically collect additional performance measures in addition to the regulatory metrics. These measures provide transit agencies objective assessment of current circumstances, past trends, existing concerns, and unmet needs. They provide perspective, understanding, and context to what has happened and what is happening within the organization. They identify how well service is being provided to their customers, the areas where improvement may be needed, and the effects of action previously taken to improve performance. In turn, this can drive the provision of efficient services to retain current passengers and attract new riders. The key management uses of a performance measurement system are service monitoring, evaluation of economic performance, management functions, internal communications, development of service design standards, communication of achievements and changes, and noting of community benefits (TRB, 2003).

This study focuses on a service monitoring performance measurement system that serves as a basis for the proposed heuristic. A structured performance measurement system can help agencies select and distill key data items in order to better understand how the network is operating and to more readily identify areas of improvement. Based on a set of performance measures that reflect the goal of the transit assignment problem, the heuristic finds a solution by constraining the metrics to the relevant thresholds. An understanding of the performance measures is pertinent before they can be used by the heuristic.

3 Stakeholders of Transit Networks

There are a number of different stakeholders who are interested in transit performance because transit service directly and indirectly affects many aspects of a community. They include the transit passengers, transit agency, community members, and the roadway agency and motorists. Transit passengers are the users of transit service and are directly affected by transit performance. There are two types of passengers: choice riders and transit-dependent riders.
Choice riders have more than one travel option available to them. They can decide whether to use transit for their trips, given the attractiveness of the mode. Transit-dependent riders rely on transit to meet basic mobility needs as other modes are not as readily available to them. They are also referred to captive riders. Choice riders are the passengers most interested in performance measures because it reflects the attractiveness of transit mode. Transit agencies are the operators of transit service and have direct impact on transit performance. They make choices about how to allocate the finite amount of resources to best meet the agency’s goals and objectives, ensuring that the transit network is operating efficiently. They would also report on transit performance to other agencies providing funding support. Community members are the stakeholders, those who may directly support transit service through taxes and who may indirectly benefit from the role that transit plays in the community. Roadway agencies and motorists have some influence on transit performance because personal vehicles must share the road network with surface transit vehicles. Though roadway agencies have their own set of stakeholders, goal and objectives, they should partner with transit agencies to implement roadway infrastructure improvements that can benefit both transit and motorists. Motorists directly interact with transit vehicles on the road and may benefit when other motorists decide to use transit, reducing the number of vehicles on the road network. Each of the stakeholder groups has its own set of interests and priorities and often, these points of view can overlap (TRB, 2013).

This study focuses on two major stakeholder groups that directly influence transit performance. They are the transit agencies who offer the transit services to the passengers and the transit passengers who use the transit service. Both transit passengers and transit agencies are concerned with all aspects of transit service provisions. However, transit agencies are more concerned with transit performance from the perspective of the transit agency as a business, while transit passengers are concerned from their experience and perception of the quality of service. Previously, transit agencies were mainly interested in economic performance and maintenance. Cost and utilization measures were the key indicators used to reflect the amount of service that a transit agency could afford to provide on a route or the system as a whole. These metrics would indirectly measure passenger satisfaction with the quality of service provided, but did not provide a comprehensive quantitative measure of the quality of service. Transit agencies soon
learned that quality of service measures that focus on passenger’s point of view are crucial to transit performance. Urban transport involves millions of individual travel decisions. When passengers have good experiences with the transit system, ridership increases and improves the transit agency as a business. As such, transit agencies have become more concerned with passengers’ perception of the overall performance of public transit systems.

4  Types of Performance Measures

Performance measures quantitatively represent the quality of service factors that are focused on the passenger point of view. They can be categorized into system-level and route-level performance measures. Route-level performance measures access the efficiency and effectiveness of individual routes within the system, while system-level performance measures indicate the transit agency’s achievement of its established goals and objectives (King County Metro Transit, 2010). Both types of measures are important in order to gain a more complete picture of how well the system performs and to identify and evaluate adjustments to the system over time.

4.1  Route-Level Performance Measures

Route-level performance measures access the efficiency and effectiveness of routes within the system. They evaluate in detail the comfort and convenience aspects of the service being provided. They are typically used in comprehensive operational analyzes to describe the terms of passenger experiences and potential transit agency issues, to compare the results to established standards, and to compare changes in results from the previous analysis. If improvement is needed, transit agencies will seek to take further action to adjust the route as resources permit. This cycle is an iterative process with targets that change with each evaluation, since the performance of an individual route is compared to the performance of a group of similar routes. Key route-level performance measures include passenger loads, reliability, and travel time.
4.1.1 Passenger Loads

Passenger load measures assess whether transit agencies’ loading standards are being exceeded and if there is sufficient capacity to absorb anticipated future growth (TRB, 2013). It affects the passenger comfort of the on-board vehicle portion of a transit trip in terms of being able to find a seat and in overall crowding levels within the vehicle, which is related to the quality of service. If overcrowding on transit vehicles is found, transit agencies may need to increase service frequency or vehicle size to reduce crowding and increase passenger comfort. Depending on the type of transit vehicle used on the subject route, the performance measures used to measure passenger load quality of service can change.

For transit vehicles designed for mostly seated passengers, passenger load is typically measured using a load factor defined as passengers per seat (TRB, 2013). These vehicles, such as buses and commuter rail, have seats provided for half or more of the vehicles’ design load. Each load factor is associated with a specific level of service (LOS) letter that standardizes the crowding level descriptions of transit. Table 1 shows the LOS provided at different load factors for vehicles designed for mostly seated passengers. The load factors represent passenger load as a factor of the total seat capacity. In other words, 0.50 means that the total number of passengers in the vehicle is equivalent to 50% of the seat capacity.

Table 1. Levels of Service at Different Load Factors (TRB, 2013)

<table>
<thead>
<tr>
<th>LOS</th>
<th>Load Factor</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.00-0.50</td>
<td>No passenger needs to sit next to one another</td>
</tr>
<tr>
<td>B</td>
<td>0.51-0.75</td>
<td>Passengers have some freedom in where they sit</td>
</tr>
<tr>
<td>C</td>
<td>0.76-1.00</td>
<td>All passengers can sit</td>
</tr>
<tr>
<td>D</td>
<td>1.01-1.25</td>
<td>Comfortable standee load for design</td>
</tr>
<tr>
<td>E</td>
<td>1.26-1.50</td>
<td>Maximum schedule load</td>
</tr>
<tr>
<td>F</td>
<td>&gt; 1.50</td>
<td>Crushing load conditions</td>
</tr>
</tbody>
</table>

For transit vehicles designed mostly for standing passengers, average standing passenger space is typically used to describe the level of crowding on board the vehicle (TRB, 2013). Examples of these vehicles that have more than 50% of the design load standing include light rail, heavy rail, and subway lines. Table 2 shows the LOS provided at different load factors for vehicles designed...
for mostly standing passengers. The average standing passenger space can be measured by square feet per person or square metre per person.

Table 2. Level of Service at Different Average Standing Passenger Space (TRB, 2013)

<table>
<thead>
<tr>
<th>LOS</th>
<th>Average Standing Passenger Space (m²/person)</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>&gt; 1.00</td>
<td>Passengers are able to spread out</td>
</tr>
<tr>
<td>B</td>
<td>0.50-1.00</td>
<td>Comfortable standing load with space between passengers</td>
</tr>
<tr>
<td>C</td>
<td>0.40-0.49</td>
<td>Standing load without body contact</td>
</tr>
<tr>
<td>D</td>
<td>0.30-0.39</td>
<td>Occasional body contact</td>
</tr>
<tr>
<td>E</td>
<td>0.20-0.29</td>
<td>Approaching uncomfortable conditions</td>
</tr>
<tr>
<td>F</td>
<td>&lt; 0.20</td>
<td>Crushing load conditions</td>
</tr>
</tbody>
</table>

As shown in both tables above, each passenger load measure is associated with a LOS letter. At LOS A, passengers can spread out and use empty seats to store parcels and bags, rather than carry them on their laps. At LOS B, some passengers will have to sit next to each other while others do not. At LOS C, all passengers can still sit, but they have limited choice of seats. At LOS D, some passengers will be required to stand. At LOS E, the transit vehicle is as full as passengers will normally tolerate. And at LOS F, it is at crush loading levels in which passengers will have to wait for the next vehicle.

4.1.2 Reliability

Reliability measures assess the effects of all potential sources of delay and unreliability and reflect the reliability aspects that are under the control of the transit operator. Examples of these measures include excess wait time, missed trips, percent of schedules time in operations, and distance travelled between mechanical breakdowns. The two most widely used reliability measures in the transit industry are on-time performance and headway adherence.

On-time performance measures the arrival time of transit vehicles at the route terminal relative to the schedule, indicating the impact on schedule recovery time and the route cycle time. It applies to any transit service that operates according to a published timetable and at longer than 10 minute headways. It is less effective for shorter headways because passengers may not notice a
transit vehicle being off-schedule when transit vehicles are arriving at a high frequency. On-time performance is highly dependent on the definition of “on time,” which can vary by geographic areas of transit agencies. In the United States, buses that were more than five minutes late, were considered on time. While in Canada, “on time” is defined to be no more than three to four minutes late. To work toward a common industry definition, “on time” is defined as a departure from a time point that can range from being one minute early to five minutes late in the TCQSM (TRB, 2013). On-time performance measure is represented by the percentage of transit vehicles that departing from a time point that ranges from one minute early to five minutes late.

On the other hand, headway adherence measures the regularity of transit vehicle arrivals with respect to the scheduled headway. It also has an impact on schedule recovery time and the route cycle time and is applied to any transit vehicles operating at headways of 10 minutes or less. It is particularly used to measure bunching effects of transit vehicles on surface streets where two or more vehicles on the same route arrive together or in close succession and is then followed by a long gap between vehicles (TRB, 2013). Usually, the lead vehicle is overcrowded, having picked up its own passengers and passengers arriving early for the next service. Passengers that arrive during the gap in service would experience a longer waiting time than expected. Also, capacity would be wasted on underutilized trailing vehicles. More time would be needed at the end of the route for schedule recovery, potentially increasing the operating cost. Headway adherence is a statistical measure where the standard deviation of headways, which represents the range of actual headways, is divided by the average headway. This is known as the coefficient of variation of headways.

4.1.3 Travel Time

Travel time measures assess the operation efficiency of transit networks at the system- or route-level by measuring how long it takes to make a trip by transit. Specifically, it assesses how quickly persons or transit vehicles can travel between two points, how many transfers are required, and how variable travel times are from day to day. Transit travel time and operating speed influence service attractiveness, costs, and efficiency. Time-related measures are useful for evaluating the service quality of particular trips, while speed-related measures are useful for
evaluating the service quality along particular facilities. For example, average speed is typically used to compare peer routes or peer transit agencies. Both types of measures are useful for demonstrating the effects of traffic congestion on scheduled run times for routes.

It is important to distinguish between two types of travel time: actual travel time and perceived travel time. Actual travel time is the measured time it took a passenger to commute from point A to point B, while the perceived travel time is the passenger's observed time going from point A to point B. Depending on the comfort level experienced by the passenger, the perceived travel time can be much greater than the actual travel time. People perceive the time spent in crowded conditions to be more onerous than time spent in uncrowded conditions. Table 3 shows the perceived travel time factors applied to the actual travel time for seated passengers and standees, with respect to the comfort levels expressed as load factor LOS.

Table 3. Perceived Travel Time Factors for Siting and Standing Passengers (TRB, 2013)

<table>
<thead>
<tr>
<th>LOS</th>
<th>Description</th>
<th>Perceived Travel Time Factor</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Sitting</td>
</tr>
<tr>
<td>A</td>
<td>No passenger needs to sit next to one another</td>
<td>1.00</td>
</tr>
<tr>
<td>B</td>
<td>Passengers have some freedom in where they sit</td>
<td>1.00</td>
</tr>
<tr>
<td>C</td>
<td>All passengers can sit</td>
<td>1.10</td>
</tr>
<tr>
<td>D</td>
<td>Comfortable standee load for design</td>
<td>1.25</td>
</tr>
<tr>
<td>E</td>
<td>Maximum schedule load</td>
<td>1.40</td>
</tr>
<tr>
<td>F</td>
<td>Crushing load conditions</td>
<td>&gt; 1.40</td>
</tr>
</tbody>
</table>

Travel time analysis has conventionally used the actual travel time measure. However, perceived travel time measures have recently gained attention as it better reflect passenger's perspective on the efficiency of transit network operations with respect to travel time.

An important quality of service measure related to travel time is transit-auto travel time ratio. This ratio indicates how much longer the trip will take in comparison with the automobile, influencing transit passenger's choice to use transit. It is calculated by dividing the in-vehicle transit travel time by the in-vehicle single-occupancy auto travel time for a given trip. It can be used to evaluate a route segment, an entire transit route, or an origin-destination trip. This measure is sensitive to both route and trip speed and directness. Table 4 describes the LOS
provided at different transit-auto travel time ratios from the passenger's perspective. As the ratio increases, transit becomes less attractive for passengers.

Table 4. Level of Service at Different Transit-Auto Travel Time Ratios (TRB, 2013)

<table>
<thead>
<tr>
<th>LOS</th>
<th>Transit-Auto Travel Time Ratio</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>≤ 1.00</td>
<td>Faster trip by transit than by auto</td>
</tr>
<tr>
<td>B</td>
<td>&gt; 1.00-1.25</td>
<td>Comparable in-vehicle travel times by transit and auto</td>
</tr>
<tr>
<td>C</td>
<td>&gt; 1.25-1.50</td>
<td>Tolerable for choice riders</td>
</tr>
<tr>
<td>D</td>
<td>&gt; 1.50-1.75</td>
<td>Route trip up to 1 hr longer by transit for 40-min one-way trip</td>
</tr>
<tr>
<td>E</td>
<td>&gt; 1.75-2.00</td>
<td>Trip takes up to twice as long by transit than by auto</td>
</tr>
<tr>
<td>F</td>
<td>&gt; 2.00</td>
<td>Tedious for all riders</td>
</tr>
</tbody>
</table>

Many of these route-level performance measures can be aggregated to create system-level performance measures to evaluate the efficiency of the overall transit network operations.

4.2 System-Level Performance Measures

System-level performance measures assess how well a transit agency is meeting its goals and objectives. They are typically the aggregation of route-level measures over the entire transit network. If improvements are needed on the system level, the transit agency will seek to take further or different actions, or to change the standards and guidelines as resources permit. These system-level performance measures can be compared to thresholds or acceptable levels of performance based on how well other transit agencies are doing. They can also be used to track trends in transit provision across the nation, state, or province. Examples of measures include average system peak-period headway, system service span, average system speed, and total network travel time.
5 The SORTAH Performance Measurement System

Measuring the performance of a transit system is the first step toward efficient and proactive management. In turn, the development of structured performance measurement systems for transportation planning and operations has gained a great deal of attention recently, particularly as transportation agencies are required to provide service with diminishing resources. The purpose of the study is to develop a transit assignment heuristic that minimizes the overall congestion effects of the transit network. The heuristic is known as the System Optimal Re-routing Transit Assignment Heuristic (SORTAH), which uses a service monitoring performance measurement system as a basis for re-routing passengers to solve the transit assignment problem. The system has components of route-level and system-level performance measures when re-routing passengers and evaluating the network solution.

5.1 Goals and Objectives

The service monitoring performance measurement system was developed to serve as a framework for re-routing passengers in the transit assignment problem. The objective is to use performance measures as evaluation tools during the heuristic procedure to re-route passengers such that transit congestion is minimized. Measures are also used after the re-routing process to evaluate overall congestion on transit networks and overcrowding on transit vehicles that causes discomfort and deteriorates the service of transit networks. Performance measures related to service delivery, capacity and travel time of transit services would play a role in minimizing the overall congestion of transit networks.

5.2 The Performance Measures

This performance measurement system consists of the following performance measures to re-route passengers and evaluate the operational efficiency of the overall network:

- Load Factor – used during the re-routing process to reflect passenger comfort and crowding levels within the transit vehicle
• Measure of Unfairness – used during the re-routing process to reflect passenger delay on the same OD pair in an attempt to minimize total travel time

• Measures of Overcrowding – used after the re-routing process to measure operational efficiency with respect to overall comfort and crowding levels on the network

• Perceived Travel Time Measures – used after the re-routing process to measure operational efficiency with respect to travel time on the network

This section discusses the significance of each performance measure and its relation to minimizing transit congestion.

5.2.1 Load Factor

Load factor is the ratio of utilized to offered capacities on a route segment, which is also known as the capacity utilization coefficient (Vuchic, 2005). The offered capacity typically includes both standing and sitting capacity, but in the TCQSM, only the offered seat capacity is used. The load factor is slightly modified where the total demand is divided by the offered seat capacity on that route segment (TRB, 2013). The load factor is represented as follows:

\[ \alpha = \frac{P}{C} \]

where
\( \alpha \) = load factor (persons/seat)
\( P \) = flow of passengers per hour that are actually transported or utilized capacity (persons/hour)
\( C \) = offered seat capacity (seats/hour)

As discussed previously, load factor assesses the quality of service that reflects passenger comfort of the on-board vehicle portion of a transit trip in terms of being able to find a seat and in overall crowding levels within the vehicle. The quality of service is represented by a load factor LOS rating. This measure is typically used for transit vehicles designed for mostly seated passengers, in which seats are provided for half or more of the vehicles’ design load (TRB, 2013). However, it can still be applied to transit vehicles designed mostly for standing
passengers. Transit vehicles designs are reflective of the transit mode it serves. Depending on the transit mode, passengers would set different expectations on the loading standard of their trip. In turn, the load factor LOS standard can change with respect to the type of transit mode.

There are five types of transit modes: bus transit, demand responsive transit (DRT), vanpool, rail transit, and ferry transit. Bus transit is the most commonly used form of public transport in North America, because it is highly flexible. Service can be provided by many different types of vehicles, operated on many different types of right-of-way, and implemented with a variety of stopping patterns. They typically provide fixed-route service along designated routes and operate at set times or headways. In 2011, 56% of transit trips on large Canadian transit systems were by bus mode (TRB, 2013). DRT is a user-oriented form of public transportation. It is characterized by flexible routing and scheduling of small to medium-size vehicles operating in shared-ride mode between pick-up and drop-off locations according to passengers' needs. Vanpool is similar to DRT, but shared rides are provided in vans or buses between homes or a central location to a regular destination. Rail transit is similar to bus transit, but is guided by specifically designed rail tracks for rail transit submodes such as heavy rail, light rail, commuter rail and automated guideway transit (AGT). These rail tracks outline the designated routes for transit and provide fixed-route service by operating at set times or headways. In most cases, rail transit operates on grade-separated tracks while some operate on exclusive tracks that are subject to external traffic interference. Ferry transit provides a water connection between or among points where land routes are interrupted by water. Pedestrian, cyclists, and vehicles are transported across waterways where transportation connections are desirable but conditions do not justify a bridge or tunnel. Of the five modes described, bus and rail are the most common types of transit modes.

Most transit network systems mainly consist of bus transit and rail transit. For example, Toronto and Montréal are the two most populous metropolitan areas in Canada that are served by rich networks of transit lines and routes. TTC operates the Toronto transit system, which consists of 141 bus routes, 11 streetcar routes, three subway lines, and one intermediate capacity transit system as of 2012 (TTC, 2013). On the other hand, STM operates the Montréal transit system, which consists of four subway lines and 219 bus routes as of 2012 (STM, 2012). Both systems
primarily consist of bus and rail transit. Thus, the development of the load factor LOS standards will focus on bus transit and rail transit.

As per TCQSM, levels of service ratings range from LOS A (least critical rating) to LOS F (most critical rating). Each LOS rating is assigned a load factor range, representing passenger load as a factor of the seat capacity (TRB, 2013). Depending on the transit mode, the load factor ranges can vary for LOS D to LOS F. At these LOS conditions, passengers experience some crowding to overcrowding on the transit vehicles. Load factor ranges can differ because passengers have better tolerance of crowding levels for vehicles that are designed for mostly standing passengers (i.e. subway) compared to vehicles designed for mostly seated passengers (i.e. buses). On the other hand, the load factor ranges will remain consistent for all transit modes from LOS A to LOS C. These LOS ratings describe non-crowding conditions of transit vehicles, where at most, every passenger has a seat. There is no crowding because there is still a lot of room for an additional passenger who has to stand. Table 5 outlines the load factor LOS standards that are consistent for all transit modes, ranging from LOS A to LOS C.

**Table 5. Load Factor LOS A to LOS C Standards for All Transit Modes (TRB, 2013)**

<table>
<thead>
<tr>
<th>LOS</th>
<th>Load Factor</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.00-0.50</td>
<td>No passenger needs to sit next to one another</td>
</tr>
<tr>
<td>B</td>
<td>0.51-0.75</td>
<td>Passengers have some freedom in where they sit</td>
</tr>
<tr>
<td>C</td>
<td>0.76-1.00</td>
<td>All passengers can sit</td>
</tr>
</tbody>
</table>

To determine the load factor LOS standards for LOS D to LOS F, typical ranges of vehicles seat capacities and crush capacities of different transit modes are required to determine the crush load factor. The crush load factor defines the lower bound of LOS F and is calculated as follows:

\[
a_{\text{crush}} = \frac{P_{\text{crush}}}{C}
\]

where

- \(a_{\text{crush}}\) = crush load factor (persons/seat)
- \(P_{\text{crush}}\) = crush capacity (persons/hour)
- \(C\) = seat capacity (seats/hour)
Table 6 outlines the typical ranges of vehicles seat capacities and crush capacities of bus and rail transit modes (Qin, 2014), as well as the calculated crush load factor ranges for each transit submode.

Table 6. Typical Seat and Crush Capacities of Bus and Rail Transit Modes

<table>
<thead>
<tr>
<th>Mode</th>
<th>Transit Submode</th>
<th>Seat Capacity</th>
<th>Crush Capacity</th>
<th>Crush Load Factor</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bus</td>
<td>Single-decker</td>
<td>35-50</td>
<td>43</td>
<td>55-80 68</td>
</tr>
<tr>
<td>Bus</td>
<td>Double-decker</td>
<td>60-75</td>
<td>65</td>
<td>90-100 95</td>
</tr>
<tr>
<td>Bus</td>
<td>Articulated</td>
<td>60-80</td>
<td>70</td>
<td>100-160 130</td>
</tr>
<tr>
<td>Rail</td>
<td>Tram/Streetcar</td>
<td>40-60</td>
<td>50</td>
<td>90-150 120</td>
</tr>
<tr>
<td>Rail</td>
<td>LRT</td>
<td>60-80</td>
<td>70</td>
<td>160-180 170</td>
</tr>
<tr>
<td>Rail</td>
<td>Subway</td>
<td>60-80</td>
<td>70</td>
<td>200-300 250</td>
</tr>
</tbody>
</table>

As shown in Table 6, the crush load factors for bus transit ranges from 1.33 to 2.00. This range is comparable to the bus TCQSM standard crush load factor of 1.50 for LOS F (TRB, 2013), showing the load factor standard is empirically supported and should be used for all bus modes. For rail modes, the crush load factor ranged from 2.25 to 3.75. This is a large range and therefore needs further analysis. Light rail submodes that include streetcar and LRT have crush load factors ranging from 2.25 to 2.67. This is comparably the same crush load factor. For heavy rail submodes such as subway, the crush load factor is 3.33-3.75. This range is comparable to the subway TCQSM standard crush load factor of 3.00 for LOS F (TRB, 1999). Thus, light rail submodes and heavy rail submodes have different load factor LOS standards.

Three load factor LOS standards have been developed for typical transit networks systems, in which each standard represents a transit mode: bus, light rail, or heavy rail. The LOS standards are described in detail in the following sections.

5.2.1.1 Bus Mode

Bus services are designed mostly for seated passengers. It is provided by vehicle types ranging from single-decker buses to articulated and double-decker buses. Historically, the 12m single-decker bus was widely used for bus services. Nowadays, the type of bus selected for bus services depend on the quality of service desired and required capacity. The relative crush load factor for
the different types of bus vehicles is consistent, because passengers have consistent expectations on the operations and crowding levels of buses. Following the TCQSM guidelines, Table 7 outlines the load factor LOS standards for all bus modes (TRB, 2013).

Table 7. Load Factor Level of Service Standard - Bus Mode

<table>
<thead>
<tr>
<th>LOS</th>
<th>Load Factor</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.00-0.50</td>
<td>No passenger needs to sit next to one another</td>
</tr>
<tr>
<td>B</td>
<td>0.51-0.75</td>
<td>Passengers have some freedom in where they sit</td>
</tr>
<tr>
<td>C</td>
<td>0.76-1.00</td>
<td>All passengers can sit</td>
</tr>
<tr>
<td>D</td>
<td>1.01-1.25</td>
<td>Comfortable standee load for design</td>
</tr>
<tr>
<td>E</td>
<td>1.26-1.50</td>
<td>Maximum schedule load</td>
</tr>
<tr>
<td>F</td>
<td>&gt; 1.50</td>
<td>Crushing load conditions</td>
</tr>
</tbody>
</table>

5.2.1.2 Light Rail Mode

Light rail modes are designed for mostly standing passengers. It is characterized by its versatility of operation. It can operate separated from other traffic below grade, at-grade, or on an elevated structure, or operate together with motor vehicles on the surface. Single unit or multiple-unit trains can be used to service the passengers. Light rail modes include LRT and streetcars. LRT and streetcar are similar but different in many ways. They both operate with relatively frequent service. However, LRT is designed to serve fairly wide-spaced stations outside the downtown core and closely-spaced stops in downtown, while streetcars are only designed to service closely-spaced stops in downtown (Walker, 2010). The design of streetcar operations date back further in time than LRT; LRT began as an evolutionary development of streetcars to allow high speeds and provide good access to people who want to use its higher-speed segments (TRB, 2013). Table 8 outlines the load factor LOS standards for light rail mode such as LRT and streetcars. The minimum of the range was taken for the crush load factor threshold to capture the all possible crush loads of light rail mode.
Table 8. Load Factor Level of Service Standard - Light Rail Mode

<table>
<thead>
<tr>
<th>LOS</th>
<th>Load Factor</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.00-0.50</td>
<td>No passenger needs to sit next to one another</td>
</tr>
<tr>
<td>B</td>
<td>0.51-0.75</td>
<td>Passengers have some freedom in where they sit</td>
</tr>
<tr>
<td>C</td>
<td>0.76-1.00</td>
<td>All passengers can sit</td>
</tr>
<tr>
<td>D</td>
<td>1.01-1.70</td>
<td>Comfortable standee load for design</td>
</tr>
<tr>
<td>E</td>
<td>1.70-2.25</td>
<td>Maximum schedule load</td>
</tr>
<tr>
<td>F</td>
<td>&gt; 2.25</td>
<td>Crushing load conditions</td>
</tr>
</tbody>
</table>

5.2.1.3 Heavy Rail Mode

Heavy rail modes are designed for mostly standing passengers. They are known as subway, elevated rapid transit, metro, and rapid rail. They are characterized by fully grade-separated rights-of-way, high-level platforms, and high-speed multiple-unit trains. The loading and unloading of passengers at stations is rapid due to level access and multiple double-stream doors. Passengers can tolerate higher crushing load conditions because heavy rail is designed to carry high volumes of passengers. A necessary population and job density is required to support a heavy rail system, which leads it to be the dominant transit mode in the large metropolitan areas. Though the empirical data showed the crush load factor range minimum to be 3.33, the TCQSM load factor LOS guidelines with a crush load factor of 3.00 for subways is followed. Table 9 outlines the load factor LOS standards for heavy rail mode.

Table 9. Load Factor Level of Service Standard - Heavy Rail Mode

<table>
<thead>
<tr>
<th>LOS</th>
<th>Load Factor</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.00-0.50</td>
<td>No passenger needs to sit next to one another</td>
</tr>
<tr>
<td>B</td>
<td>0.51-0.75</td>
<td>Passengers have some freedom in where they sit</td>
</tr>
<tr>
<td>C</td>
<td>0.76-1.00</td>
<td>All passengers can sit</td>
</tr>
<tr>
<td>D</td>
<td>1.01-2.00</td>
<td>Comfortable standee load for design</td>
</tr>
<tr>
<td>E</td>
<td>2.00-3.00</td>
<td>Maximum schedule load</td>
</tr>
<tr>
<td>F</td>
<td>&gt; 3.00</td>
<td>Crushing load conditions</td>
</tr>
</tbody>
</table>

SORTAH uses the load factor LOS standards for the passenger re-routing process. As per TCQSM, average standing passenger space LOS standard is recommended for transit vehicles designed mostly for standing passengers. However, the amount of floor space per passenger is an
arbitrary measure and should be used to set the comfort standards (House of Commons Transport Committee, 2003). Load factor LOS standards were developed for both transit vehicles designed mostly for seated passengers and transit vehicles designed mostly for standing passing because they take into account the detail of transit vehicle type, such as the number of seats provided and the provision made for passengers to stand in comfort with adequate hand holds.

5.2.2 Measure of Unfairness

Measure of unfairness is the ratio of a passenger’s current travel time to the re-routed travel time. It assesses passenger delay on the same OD pair, assisting in the passenger re-routing process such that the system optimal solution is constrained with a reasonable set of allowable paths. It is calculated as follows:

$$\phi = \frac{t_{re-routed}}{t_{current}}$$

where

- $\phi$ = measure of unfairness
- $t_{re-routed}$ = travel time in re-routed conditions (minutes)
- $t_{current}$ = travel time in current conditions (minutes)

Conventionally, transit assignment models follow user equilibrium principle where passengers are assumed to choose a path such that the journey times are equal or less than other allowable paths not used by the passenger. In other words, passengers travelling between the same OD pair experience similar travel times defining the fairness property. On the other hand, the system optimal principle assumes travellers are allocated centrally by explicitly minimizing the total travel time. In a pure system optimal solution, passengers could be re-routed to unreasonably long paths. To honour that only a few passengers are willing to sacrifice their shortest route for the benefit of the community, measure of unfairness limits the possible re-routing paths of a passenger on the same OD pair. The measure is used to minimize marginal travel times for each passenger to reduce total travel time.
5.2.3 Measures of Overcrowding

Measures of overcrowding examine the operational efficiency of a network with respect to crowding and comfort level of a network on a segment- and route-basis. Three measures of overcrowding were developed: measure of route overcrowding, measure of route overcrowding change and measure of segment overcrowding change.

Measure of route overcrowding is the ratio of the number of routes with LOS F segments to the total number of routes in the network. It assesses the number of routes affected by congestion. It is calculated as follows:

\[ \Omega_{\text{route}} = \frac{R_{\text{LOS F}}}{R_{\text{network}}} \]

where

- \( \Omega_{\text{route}} \) = measure of route overcrowding
- \( R_{\text{network}} \) = total number of routes in the network
- \( R_{\text{LOS F}} \) = total number of routes with LOS F segments

Measure of route overcrowding change is the ratio of the difference between the initial number of routes with LOS F segments and the number of routes with LOS F segments after the re-routing process, to the initial number of routes with LOS F segments in the network. It assesses the change in the number of routes affected by congestion and how well SORTAH has reduced network congestion. It is calculated as follows:

\[ \Delta_{\text{route}} = \frac{(R_{\text{LOS F\_re-routed}} - R_{\text{LOS F\_base}})}{R_{\text{LOS F\_base}}} \]

where

- \( \Delta_{\text{route}} \) = percentage change in route overcrowding
- \( R_{\text{LOS F\_re-routed}} \) = total number of routes with LOS F segments in the re-routed network (after the final iteration of the re-routing process)
- \( R_{\text{LOS F\_base}} \) = total number of routes with LOS F segments in the base network (before the re-routing process)
Another measure to assess the effectiveness of reducing network congestion by SORTAH is the measure of segment overcrowding change. Examining the individual segments instead of the affected routes, it is the ratio of the difference between the initial number of LOS F segments and number of LOS F segments after the re-routing process, to the initial number of LOS F segments in the network. The equation is as follows:

\[
\Delta_{\text{segment}} = \frac{(N_{\text{LOS F re-routed}} - N_{\text{LOS F base}})}{N_{\text{LOS F base}}}
\]

where

\(\Delta_{\text{segment}}\) = percentage change in segment overcrowding

\(N_{\text{LOS F re-routed}}\) = number of segments with LOS F in the re-routed network (after the final iteration of the re-routing process)

\(N_{\text{LOS F base}}\) = number of segments with LOS F in the base network (before the re-routing process)

These measures of change can range from 0.00 to 1.00. A value of 0.00 indicates that SORTAH has eliminated overcrowding on the network and a value of 1.00 indicates that the amount of overcrowding before and after the re-routing process is the same. Values would not exceed 1.00 because SORTAH is designed to reduce overcrowding, even if it is only a small change. The desired value for the measures of change is less than 0.50, indicating that the original crowding level has been reduced by half.

5.2.4 Perceived Travel Time Measures

Perceived travel time measures assess the operational efficiency of a network with respect to travel time. Two measures of perceived travel time were developed: measure of total travel time change and measure of average travel time change.

Measure of total travel time change is the ratio of the difference between the initial total travel time and the total travel time after the re-routing process, to the initial total travel time of the network. It assesses the effectiveness of SORTAH in reducing network congestion with respect to the total travel time. It is calculated as follows:
\[ tt = \sum tt_i \]
\[ \Delta tt = \frac{(tt_i LOS F_{re-routed} - tt_i LOS F_{base})}{tt_i LOS F_{base}} \]

where
\( tt = \) total travel time per passenger (minutes)
\( tt_i = \) travel time perceived by each passenger (minutes)
\( \Delta tt = \) percentage change in total travel time
\( tt_i LOS F_{base} = \) total travel time for base network (before the re-routing process)
\( tt_i LOS F_{re-routed} = \) total travel time for re-routed network (after the re-routing process)

Measure of total travel time change is only applicable if demand and supply is constant. This can be used to compare the user equilibrium solution with the system optimal solution of the transit network because both solutions are based on constant supply and demand. When two networks of different supply and demand are to be compared, the normalized measure of the total travel time should to be used for the same metric base. The equation is as follows:

\[ tta = \frac{tt}{P_{total}} \]
\[ \Delta tta = \frac{(tta LOS F_{re-routed} - tta LOS F_{base})}{tta LOS F_{base}} \]

where
\( tt = \) total travel time per passenger (minutes)
\( tta = \) average travel time per passenger (minutes/person)
\( P_{total} = \) total number of passengers served by the system (persons)
\( \Delta tta = \) percentage change in average travel time
\( tta LOS F_{base} = \) average travel time for base network (before the re-routing process)
\( tta LOS F_{re-routed} = \) average travel time for re-routed network (after the re-routing process)

The total and average travel times are based on the perceived travel time of each individual using the network. Each individual’s travel time is composed of various travel time components: wait time, in-vehicle time, transfer time, and access or egress time. Wait time is the time a passenger takes standing at a stop or station waiting for the vehicle to arrive before boarding. In-vehicle time is the time a passenger spends onboard the transit vehicle traversing from boarding stop to
alighting stop. Transfer time is the time a passenger takes to transfer from one route to another, in order to arrive at his or her destination. The transfer time includes the walk time between the alighting stop of the previous route and the boarding stop of the target route, and the wait time for the latter route. Access time is the time it takes a passenger to walk from his or her origin to the origin stop, while egress time is the walk time from his or her destination stop to the destination. The method of access to the origin stop or egress from the destination stop is not always walking. It can range from walking to cycling to driving, depending on the type of system. For an urban transit system, access and egress times are associated with walking, and sometimes cycling; for a commuter rail system, they are associated with driving or cycling. Figure 6 illustrates the travel time components for a sample trip.

**Figure 6. Travel Time Components**

<table>
<thead>
<tr>
<th>Time</th>
<th>Distance</th>
</tr>
</thead>
<tbody>
<tr>
<td>ORIGIN: Departure Time</td>
<td>Origin Waiting Time</td>
</tr>
<tr>
<td>ON STOP: Transfer</td>
<td>Transfer Time</td>
</tr>
<tr>
<td>OFF STOP: Transfer</td>
<td>In-Vehicle Time</td>
</tr>
<tr>
<td>ON STOP: Transfer</td>
<td>Transfer Time</td>
</tr>
<tr>
<td>OFF STOP: Transfer</td>
<td>In-Vehicle Time</td>
</tr>
<tr>
<td>ORIGIN STOP</td>
<td>Access Time</td>
</tr>
<tr>
<td>DESTINATION STOP</td>
<td>Egress Time</td>
</tr>
<tr>
<td>DESTINATION: Arrival Time</td>
<td>In-Vehicle Time</td>
</tr>
</tbody>
</table>

**Figure 6 Diagram**

- **Access Time**
- **Origin Waiting Time**
- **Transfer Time**
- **In-Vehicle Time**
- **Egress Time**
- **In-Vehicle Time**
- **Destination Stop**
From a passenger’s perspective, each travel time component requires different physical efforts. Wait time, transfer time, and the access and egress time are perceived to be more onerous than in-vehicle time. Passengers view these stages of the trip to have little or no productivity, especially when they encounter wayfinding difficulty and experience general anxiety associated with getting to a particular location. Travel time factors need to be applied to the actual travel time components to calculate the passenger’s perceived travel time. When the perceived individual travel times are used to compute the total network travel time and the average travel time, it reflects the passenger’s perception of the quality of service.

Table 10 shows ranges of different travel time component factors relative to the in-vehicle time component, from various U.S. and U.K. studies (Pratt & Evans, 2004; Wardman, 2004; TRB, 2013). For the purpose of this study, the U.S. averages are used because Canada and U.S. cities are comparable. The walk time factor would be applied to the access or egress component.

Table 10. Relative Travel Time Component Factors (TRB, 2013)

<table>
<thead>
<tr>
<th></th>
<th>In-Vehicle Time</th>
<th>Walk Time</th>
<th>Initial Wait Time</th>
<th>Transfer Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>U.S. average</td>
<td>1.0</td>
<td>2.2</td>
<td>2.1</td>
<td>2.5</td>
</tr>
<tr>
<td>U.S. range</td>
<td>1.0</td>
<td>0.8-4.4</td>
<td>0.8-5.1</td>
<td>1.1-4.4</td>
</tr>
<tr>
<td>U.K. average</td>
<td>1.0</td>
<td>1.7</td>
<td>1.8</td>
<td>N/A</td>
</tr>
</tbody>
</table>

The perceived individual travel time is calculated as follows:

\[
\text{tt}_i = \beta_a(\text{at}) + \beta_w(\text{wt}) + \beta_v(\text{vt}) + \beta_t(\text{ft}) + \beta_e(\text{et})
\]

where

- \( \text{tt}_i \) = travel time experienced by each passenger (minutes)
- \( \text{at} \) = actual access time (minutes)
- \( \text{wt} \) = actual initial wait time (minutes)
- \( \text{vt} \) = actual in-vehicle time (minutes)
- \( \text{ft} \) = actual transfer time (minutes)
- \( \text{et} \) = actual egress time (minutes)
- \( \beta_a \) = access time factor (2.2)
- \( \beta_w \) = initial wait time factor (2.1)
- \( \beta_v \) = in-vehicle time factor (1.0)
- \( \beta_t \) = transfer time factor (2.5)
- \( \beta_e \) = egress time factor (2.2)
Chapter 4
SORTAH: System Optimal Re-routing Transit Assignment Heuristic

1 Chapter Overview

This chapter presents the conceptual development of the proposed heuristic approach. It describes the detailed structure of the System Optimal Re-routing Transit Assignment Heuristic – SORTAH.

- Section 2 outlines the basic assumptions of the proposed heuristic. It is developed to re-route passengers of a user equilibrium solution to a solution that is close to system optimum, while accounting for user constraints.

- Section 3 provides the modelling framework of SORTAH. The objective is to minimize congestion (in other words, total travel time and overcrowding) while honouring passenger travel needs.

- Section 4 describes the heuristic procedure of SORTAH. It utilises performance metrics, marginal travel times and threshold ratios to perform the re-routing procedure.

- Section 5 outlines the characteristics and limitations of SORTAH.

- Section 6 reviews the application opportunities of the transit assignment heuristic.

2 Basic Assumptions

Congestion occurs because the network supply is not able to serve for the excess travel demand during a given operational interval, which is typically during the peak periods. Congestion can be experienced by passengers when travelling in transit vehicles, waiting on the platforms, and
walking in the pedestrian channels that link platforms and stations entries or exits. Transit congestion is primarily associated with overcrowding on the public transit system. Occasional occurrence of this phenomenon may be a sign of success, but chronic overcrowding at a scale such that passengers cannot board trains would impose significant costs to the city’s businesses and economy.

Overcrowding typically occurs on specific route segments or “pinch points” in the transit network. Transit congestion occurs when one or more routes are experiencing crushing passenger loads and there is more than one “pinch point” in the transit network. “Pinch points” are defined to be route segments with a LOS E rating. These are the areas that require more attention in mitigating the congestion effects. Through identification of these problem areas, passengers can be re-routed to less utilized routes, effectively redistributing load among the network.

SORTAH is a heuristic developed to re-route passengers of a user equilibrium solution to a user-constrained system optimal solution for a congested transit network. These are networks characterized by having more than one overcrowded “pinch points” with crushing passenger loads. The objective is to relieve the network of these “pinch points” from crushing passenger loads while integrating passenger constraints to increase network efficiency and guaranteeing superior fairness compared to a pure system optimal solution. Therefore, the proposed re-routing strategies would be acceptable to the users of the transit system. The heuristic requires detailed information of each passenger’s path choice. The disaggregation of the data requirement limits SORTAH to be compatible with microsimulation outputs that provide details of passenger’s path choices, such as the path set and the associated travel time components and departure times for each path.

3 Modelling Framework

This section outlines the modelling framework of SORTAH. It uses a performance management system to drive the path allocation model that re-route passengers such that the solution performs
better than user equilibrium and simultaneously guarantees superior fairness compared to pure system optimum.

3.1 Network Representation

In SORTAH, the supply model serves as a basis for analyzing the network effects of re-routing passengers. The transit network is represented by a linked graph with elements of stops, route segments, routes, bounds, branches, vehicle types and transfer group. These terms are described below and graphically represented in Figure 7:

- **Stops** – They are fixed entities along a route, also known as nodes. They allow for transfer if the stop is accessible to another stop and allows for transfer-connection.

- **Route Segments** – They are the portion of a branch between two consecutive stops. These segments are associated with a vehicle type, that allows for load factor LOS standards to be applied to rate the overcrowding characteristic of the segment.

- **Routes** – They are a pre-defined set of nodes in which transit vehicles can traverse through following a schedule, and are directionless.

- **Bounds** – They represent one possible direction of a route. They are either “inbound” or “outbound”.

- **Branches** – They are associated with a specific route and a specific bound. They are defined by a sequence of transit stops, characterized by a set of service characteristics. A vehicle type would be associated with each branch.

- **Vehicle Types** – They are the types of transit vehicles associated with a particular type of transit mode. For example, some branches may be a bus routes and is associated with a bus vehicle, while other branches may be subways lines and is associated with subway trains. Each vehicle type has its own crushing capacity which would be used to determine the load factor LOS of the specific route segment.
- **Transfer Group** – They are a set of stops that are part of different routes that allow passengers to go from one route to another.

**Figure 7. Network Representation**

SORTAH analyzes the network at the route segment level. This approach is convenient because passengers can examine travel time on a route segment using only the initial wait time at the origin stop and the travel time from the origin stop to the destination stop upon boarding a vehicle on that route. Load factor and measure of unfairness can also be examined for each segment before considering re-routing a passenger to another path within the passenger’s path set. These path sets are developed assuming a mental model structure of individuals traversing through a travel network modelled by sets of route segments.

### 3.2 Passenger’s Path Set

In the literature, choice set generation models are rarely explicitly specified and are calibrated based on indirect information (Cascetta et al., 2002). Mathematical models would be derived to represent route choice behaviour of passengers using information such as wait times, expected travel times and the number of transfers. These mathematical models can unrealistically assume that passengers have full information of the entire system if the models are deterministic or do
not account for the explicit manipulations of a traveller’s learning and adaptation if the models are stochastic. To explicitly represent the learning and decision-making activities of each passenger, SORTAH assumes that passenger’s path sets are developed based on a mental model structure that follows a stochastic process approach. Individual travel decisions are based on accumulated experience (i.e. mental models) gathered from repetitively travelling through the transit network on consecutive days, which is the same as the concept implemented by Wahba (2005) for transit path choice and similar to the model developed by Ettema et al. (2005) for auto path choice. This assumption of the path generation model represents the user constraint aspect of SORTAH’s user-constrained system optimal solution.

The theory of mental models is widely studied in psychology (Johnson-Laird, 1983). Individuals learn through experimentation and experience collection to form their mental models. Mental models are the internal representations that individuals’ cognitive systems create to interpret their surroundings (i.e. part of the real world that is relevant to the individual). They are used as a basis for choosing an adequate action by using their beliefs and knowledge about the world to make predictions about the future and the consequences of any action. Beliefs and knowledge are formed through information that is accumulated over time, such as results of different actions that have been brought about and the experiences and information obtained from different sources. New information would be interpreted and included according to and within existing mental models. Individuals would memorize their experience with certain situations and use this experience collection to choose a rewarding action. While learning techniques are based on collecting experiences, they usually differ in the way in which experience is processed to provide expectations and predictions about the outputs of actions.

Passengers plan their trips in non-stationary and uncertain environments. By using a mental model representation for the transit network and its conditions, the dynamics of individual’s adaptation of the network effects can be modelled. Since travellers do no possess perfect knowledge of the whole transportation system, they learn and plan activities according to their stored information related to their trip such as the best (minimum) and worst (maximum) waiting and travel times for a specific run. In other words, passengers make transit trip choices
contingent upon outcomes of previous choices (i.e. experience with the transit system performance). Passenger path sets are generated through an iterative process that solves for a user equilibrium transit assignment problem. Once a steady state is reached, the user constrained path sets can be used by SORTAH to solve for the user-constrained system optimal solution. Paths in each path set are determined by following a number of dominance and heuristic rules of learning:

- The acceptable number of transfers for any path of a passenger’s OD pair is not more than one plus the minimum number of transfers for the same OD pair. Path options with two or more transfer connection greater than the minimum number of transfers for a given trip are not considered. For example, if the minimum number of transfers to complete a certain OD trip is one, then path options with a number of transfers greater or equal to three are not generated. Meanwhile, path options with two transfers for the same trip are considered.

- Access/egress distances from stops are within an acceptable access distance buffer zone for transit passengers. When the access/egress mode is walking, the maximum walking distance is 1,000 meters (Alshalalfah & Shalaby, 2007). The catchment area is expected to vary by transit mode. Passengers are willing to walk more to a subway station compared to a bus stop. The maximum walking distance is assumed as the upper bound for the rapid rail transit mode in order to mitigate any impact of the randomness in generating geo-locations for trip-ends.

- Passengers are aware of available transit routes and possible transfers from the stops they are boarding.

- The common lines problem is explicitly dealt with where passengers who choose an origin boarding stop can be faced with more than one route that traverses through the stop.
• The difference between frequent and occasional passengers of the transit network is represented by their awareness level. Frequent users have high level of awareness and their set of transit paths contain more options than the set for occasional users.

• Not all perceived experience or information is stored in the mental model to avoid modelling biased perceptions that can lead to a deluded equilibrium state (Nakayama & Kitamura, 2000). Experiences and information that are not representative of the transit conditions (i.e. severe congestion due to an accident) are ignored or stored with low reliability value.

• Generalized cost function is used to evaluate each path. They capture within-day and real-time dynamics through changing attributes that are direct functions of service features. These include waiting time, travel time, and transfer time.

• Each passenger decides on the departure time and the stop-route sequence to take for each iteration of the transit trip. Depending on the departure time and stop-route sequence, new experiences and information are perceived and integrated into the passenger’s mental model in the light of previous knowledge. This captures evolution of day-to-day experiences and conditions.

• The final passenger path set generated by the mental model has no more than seven paths. This is because passengers realistically consider between four to seven alternatives for travel choices (Cascetta et al., 2002).

Figure 8 illustrates a passenger path set used in SORTAH. Each path is represented by its origin, series of stops and routes, and destination. Each stop and route element of a path is described by its travel time components including access time, initial wait time, in-vehicle times, transfer times, and egress time. A departure time and the number of transfers are also used to describe each path.
Figure 8. Passenger Path Set

Each Path:
- Travel Time Components \{Access Time, Initial Wait Time, In-Vehicle Time(s), Transfer Time(s), Egress Time\}
- Stop-Route Sequence Components \{O-Stop, Route1, Transfer Stop(s) [Off Stop(s), On Stop(s)], ..., RouteN, D-Stop\}
- \# of Transfers \{N-1\}
- Departure Times \{Path Dependent\}
- **Majority Path:** A path that passenger takes in the user equilibrium solution

Using the passenger’s path set, SORTAH re-routes passengers away from the path that passengers use most during the iterative process. This is known as the *majority path* and the path that passengers take in the user equilibrium transit assignment solution. In SORTAH, re-routing continues accounts for the within-day dynamics of a network. Network effects influence travel time of each path and therefore travel time components are updated each time a passenger is re-routed. However, the day-to-day dynamics are relaxed. Mental models are not applicable during the re-routing process because system optimal solutions assume people do not have the choice to take a route according to their experience. Passengers are assigned to a path following the path allocation model of SORTAH. The use of mental models in the development of passenger path sets is crucial for the re-routing process, because passengers are more receptive of being re-routed to another path they are familiar with and have learned through experience. Bounding the set of paths in which passengers can take through their mental model accounts for the user constraint component of the user-constrained system optimal solution. The re-routing of
passengers, which follows the path allocation model, would be more acceptable by the users of the system.

3.3 Path Allocation Model

Traditionally, path choice models were the fundamental core of transit assignment models. They modelled how passengers chose their paths to complete their journey. Passengers were assumed to choose a cost-efficient path to complete his or her journey following Wardrop’s user equilibrium principle:

“the journey times on all the routes actually used are equal, and less than those which would be experienced by a (passenger) on any unused route.”

However, SORTAH does not follow this principle. Rather, passengers are allocated centrally to minimize the total cost incurred by all passengers, following Wardrop’s system optimal principles:

“the average journey time is a minimum.”

While following this principle, SORTAH also considers individual travel needs by imposing additional constraints to ensure passengers are assigned to acceptable paths. Thus, a path allocation model was developed.

The path allocation model is the fundamental core of SORTAH. It uses the service monitoring performance measurement system to model how passengers are centrally allocated among the possible paths to complete their journey. Conventionally, system optimal models primarily minimize total travel time to obtain the system optimal solution. Crowding levels are typically neglected or not explicitly considered when solving for a system optimal solution. The most effective algorithm for re-routing passengers accounts for both crowding level and travel time factors when re-routing passengers. SORTAH’s objective function is to minimize crowding levels and measures of unfairness to achieve the user-constrained system optimal solution. This model reacts to the changing transit performance in an attempt to improve services by reducing
crowding and total travel time. Re-routing of passengers follow a set of criteria, and is then evaluated for its effectiveness.

3.3.1 Re-Routing Criteria

There are four main criteria for re-routing passengers:

1. Path Set Criterion: Passengers must be re-routed to another path that is within the passenger’s path set.

2. Current Crowding Level Criterion: Passengers must be re-routed to another path if the current path he or she is traversing has a LOS F route segment.

3. Target Crowding Level Criterion: Passengers must be re-routed to a target path that does not have a route segment with a load factor LOS F.

4. Target Travel Time Criterion: Passengers must be re-routed to a target path that has the smallest measure of fairness.

The first two criteria (i.e. path set criterion and current crowding level criterion) must simultaneously be met in order for re-routing of a passenger to be considered. These constraints are related to transit mode-based load factors. Route segment load factors are calculated and compared to the standards developed in the performance measurement system. Depending on the transit mode the route segments are operating in, each segment is labeled with a load factor LOS accordingly. The load factor LOS standards range from LOS A (least critical rating) to LOS F (most critical rating). These LOS ratings allow for apples-to-apples comparison of the route segments across all types of transit modes. The purpose of setting LOS F as the threshold is to minimize the crushing load conditions on the network. As discussed, crushing load conditions cause distress for passengers travelling on the transit system and inefficient operation for the transit system. Transit systems should operate at most with LOS E, which indicates a system operating with the maximum schedule load. Passengers would experience crowding, but this is considered tolerable for passengers as these crowding levels are expected during peak periods. Transit agencies design the maximum schedule load to accommodate the sudden surge of
passenger during peak periods. When transit systems are designed such that supply consistently exceeds demand, the benefits do not justify the cost. These systems are considered inefficient and very costly for transit agencies to operate. Thus, a transit network that operates at LOS E is an indication of a successful public transport system.

When a transit network operating at LOS E is about to reach its crush capacity, it is an indication to transit agencies that additional capacity may be required. Networks that operate at crushing loads are inefficient. A transit network that already has route segments with crushing loads (i.e. LOS F), attempts should be made to re-route passengers away from these segments. If the attempted re-routing method still results in a network with LOS F route segments, it is an indication that additional capacity is needed. Transit agencies should consider adding more capacity through additional service provisions or constructing transit infrastructure.

Once the consideration for re-routing a passenger is identified, the target path needs to be determined. The last two criteria are related to which target path the passenger can be re-routed to. The third criterion (i.e. target crowding level criterion) re-routes passengers to a path with respect to crowding levels. If no target paths meet the third criterion, re-routing would not occur. In other words, passengers cannot be re-routed to any paths that have a route segment with LOS F. Passengers can only be re-routed to any of the target paths that do not have LOS F segments. If there are more than one target paths with no LOS F segments, passengers would be re-routed to the path that meets the last criterion (i.e. target travel time criterion).

Prototypes of SORTAH were developed to test for variations of the target travel time criterion. It was found that the most effective algorithm for re-routing passengers accounts for both crowding level and travel time factors when re-routing passengers. The target crowding level criterion and target travel time criterion of the re-routing criteria reflect SORTAH’s objective function to minimize crowding levels and measures of unfairness to achieve the user-constrained system optimal solution. The last criterion re-routes passengers to a path with the smallest measure of unfairness. This measure is the ratio of the passenger’s current travel time to the target path’s re-routed travel time. The target path’s re-routed travel time would be altered from its original travel time to account for more dwell time required to include an additional passenger. A
measure of unfairness of below 1.0 would indicate that the travel time of the re-routed path is smaller than the current path and the total network travel time would be reduced. If the measure of unfairness is above 1.0, this would indicate that the travel time of the re-routed path is larger than the current path. This would increase the total network travel time, but the additional travel time would be minimized. The net effect should result in a reduced total network travel time for the constrained system optimal solution when comparing to the user equilibrium solution. After passengers are re-routed according to the criteria outlined, the affected paths’ travel times are adjusted.

### 3.3.2 Travel Time Adjustment

During the re-routing process, passengers that have been re-routed would alter the travel time of his or her original and new path. With one less passenger on the original path, travel time for that path would decrease because dwell time does not need to accommodate the additional person boarding and alighting the transit vehicle. On the other hand, the new path travel time would increase because dwell time needs to accommodate the additional passenger.

Dwell time is the time spent at a stop or station serving passenger movements. This is typically included in the in-vehicle time component as passengers are waiting on board for the transit vehicle to start moving. Dwell time is associated with passenger boarding and alighting volumes, vehicle type and size, and in-vehicle circulation (TRB, 2013). Dwell time increases as passenger volume increases, because transit vehicles need to spend more time at a stop for passengers to board and alight the vehicle. In addition, when transit vehicles are crowded, there are more standees and this causes more time required for passengers to circulate within the vehicle to reach the exit door. Depending on the type of transit vehicle and transit mode, the boarding and alighting time would vary. Table 11 outlines the typical ranges of boarding and alighting times for one passenger, depending on the type of transit vehicle.
Table 11. Transit Unit Standing Times for Various Transit Vehicles (Vuchic, 2005)

<table>
<thead>
<tr>
<th>Transit Vehicle</th>
<th>No. of Doors</th>
<th>Fare type/collection</th>
<th>Boarding Time</th>
<th>Alighting Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bus</td>
<td>2</td>
<td>At boarding</td>
<td>3.0-5.0</td>
<td>1.5-3.0</td>
</tr>
<tr>
<td>Bus/Streetcar</td>
<td>4</td>
<td>Flat passes, validation</td>
<td>2.0-3.0</td>
<td>1.5-3.0</td>
</tr>
<tr>
<td>Bus/Streetcar</td>
<td>4-6</td>
<td>Self-service, validation</td>
<td>1.2-2.2</td>
<td>1.0-1.6</td>
</tr>
<tr>
<td>Articulated Bus</td>
<td>8</td>
<td>Self-service, validation</td>
<td>1.6-2.6</td>
<td>1.5-3.0</td>
</tr>
<tr>
<td>LRT</td>
<td>16-24</td>
<td>Any type, prepaid</td>
<td>0.8-1.6</td>
<td>0.6-1.4</td>
</tr>
<tr>
<td>Subway</td>
<td>24-80</td>
<td>Any type, prepaid</td>
<td>0.6-1.4</td>
<td>0.5-1.2</td>
</tr>
</tbody>
</table>

On average, the boarding and alighting times would increase by 2.0 seconds for each additional passenger taking buses and streetcars, while the boarding and alighting times would increase by 1.0 second for each additional passenger taking LRT or subway. Given that the passenger paths can use a variety of transit modes, a standard 2.0 second dwell time adjustment is applied to the path travel times when passengers are re-routed. This accounts for 1.0 second of boarding and 1.0 second of alighting the transit vehicle. If the passenger is taking a path with transfers, the time adjustment would equal to the standard 2.0 second dwell time plus the number of transfers multiplied by 2.0 second dwell time. This adjustment accounts for each transit vehicle the passenger boards and alights when transferring and is calculated as follows:

\[
\begin{align*}
    tt_{\text{new, original path}} &= tt_{\text{old, original path}} - 2.0 - 2.0(nt_{\text{original path}}) \\
    tt_{\text{new, target path}} &= tt_{\text{old, target path}} + 2.0 + 2.0(nt_{\text{target path}})
\end{align*}
\]

where
- \( tt_{\text{new, original path}} \) = travel time of the original path after re-routing (minutes)
- \( tt_{\text{old, original path}} \) = travel time of the original path before re-routing (minutes)
- \( nt_{\text{original path}} \) = number of transfers of the original path
- \( tt_{\text{new, new path}} \) = travel time of the target path after re-routing (minutes)
- \( tt_{\text{old, new path}} \) = travel time of the target path before re-routing (minutes)
- \( nt_{\text{target path}} \) = number of transfers of the target path

Dwell time adjustments are applied on both the original path the passenger was taking and the new path the passenger was re-routing to. These modified individual travel times would have an impact on the total network travel time, affecting the re-routing evaluation.
3.3.3 Re-Routing Evaluation

Re-routing evaluation is used to determine whether a steady state solution has been found and to evaluate the user-constrained system optimal solution against the user equilibrium solution. There are two points during the heuristic in which the re-routing evaluation can occur:

1. After each iteration – To determine whether steady state solution has been reached by comparing the number of LOS F segments in the current iteration with the previous iteration.

2. After the final re-routing process – To compare the user-constrained system optimal solution with the user equilibrium solution.

The first type of re-routing evaluation occurs after the initial run of the passenger re-routing process. If the number of LOS F segments in the current iteration is different from the previous iteration then another incidence of the passenger re-routing process is performed. Several iterations of the passenger re-routing process is executed until the number of LOS F segments has reached a steady state (i.e. the number of LOS F segments in the current process is the same as the number of LOS F segments in the previous process). The second type of re-routing evaluation occurs after the final run of the passenger re-routing process. Using the measures of overcrowding and the perceived travel time measures, the user-constrained system optimal solution is compared to the user equilibrium solution. These measures examine how operational efficiency has improved from the user equilibrium solution to the user-constrained system optimal solution with respect to crowding levels and total travel time. It allows for the evaluation of the effects of the reduction of overcrowding on the change in total travel time.

4 The Heuristic Procedure

Prior to executing the SORTAH procedure, the user equilibrium solution needs to be represented by a detailed account of each passenger’s path set and the network’s route segments. Using the user equilibrium solution as an input, the SORTAH procedure can be applied to find the user-
constrained system optimal solution. The process assumes the four re-routing criteria, the travel time adjustment on affected paths, and the re-routing evaluation. There are three main subprocesses:

1. Pre-Process – Prepares the passenger path sets and route network data for re-routing.

2. Passenger Re-Routing Process – Re-routes passengers throughout the network using the re-routing criteria.

3. Number of Iterations – Determines the number of iterations of passenger re-routing process using the first type of re-routing evaluation. After the final iteration of the re-routing process, the second type of re-routing evaluation is applied.

Figure 9 illustrates the SORTAH procedure in a graphical form and the following outlines the heuristic steps:

Step 1.  
*Calculate the total network travel time.*  
*Calculate load factors of each route segment.*  
*Assign a LOS rating to each route segment by following the transit mode-based LOS standards.*  
*Find all routes segments with LOS F and add them to the LOS F segments list.*

Step 2.  
*Determine all passengers that take a path that traverses through a route segment with LOS F and add them to the passenger list.*

Step 3.  
*For each passenger, consider the passenger for re-routing if the passenger has another path in their path set that does not traverse through a route segment with LOS F.*  
*If there is only one other path that does not traverse through a route segment with LOS F, re-route the passenger to the target path and go to Step 5. Otherwise, go to Step 4.*

Step 4.  
*Determine the subset of target paths that the passenger can be re-routed to.*  
*Calculate the measure of unfairness for each target path.*  
*Re-route the passenger to the target path with the smallest measure of unfairness.*  
*Remove the passenger from the passenger list and go to Step 5.*
Step 5.
Adjust the travel time for the original path and the target path the passenger was re-routed to. Adjust the load factor for all segments associated with the original path and the target path, and updated the LOS rating accordingly.
Update the LOS F segments list.
Go to Step 6 if the passenger list is empty. Otherwise, go to Step 3.

Step 6.
Calculate the total network travel time and the change from the previous network travel time. Calculate the number of route segments with LOS F.
If the current iteration’s number of route segments with LOS F (#LOS_Fi) is equal the previous iteration’s number of route segments with LOS F (#LOS_Fi-1), STOP. Otherwise, go to Step 2.

Figure 9. The SORTAH Procedure
5 Characteristics and Limitations

5.1 Characteristics

SORTAH has been designed with the following characteristics that are different from conventional models:

- The heuristic produces a user-constrained system optimal solution. This solution addresses the system perspective of an overcrowded transit network while integrating individual travel needs. The resulting network would systematically perform better than a user equilibrium model, while simultaneously guaranteeing superior fairness compared to the pure system optimal model. SORTAH was developed with consideration for its applicability in the real-world and accounting for passenger travel behaviour.

- The SORTAH algorithm analyzes the transit network on a segment-basis. Crowding levels are analyzed for each segment of a route between two consecutive stops to identify “pinch points” that need to be relieved in the network. The analysis is conducted at a disaggregated level to provide insight on network effects due to individual passenger behaviour.

- The SORTAH algorithm produces results within two hours for a real-world application. The purpose of a heuristic is to quickly find a good solution to assist planners in making a well-informed decision on a policy in a timely manner. The execution time within two hours is desirable to serve as a tool for short-term decision making for planners and policy makers.

- System optimal models are conventionally solved by minimizing the total travel time, which does not necessarily reduce crowding levels on the network. Since SORTAH is designed to solve for the user-constrained system optimal network, passenger travel experience needs to be modelled directly. Crowding levels and comfort is a significant factor to passenger travel experience. Rather than minimizing the total travel time, the objective function of SORTAH is to solve for a network in which there are no more
“pinch points” or overcrowded routes in the network and minimizing the marginal travel time. Minimizing crowding levels improves passenger’s travel experience and in turn improves system-wide operational efficiency of the transit network.

- SORTAH accounts for the probabilistic effects of traveller choice. The user equilibrium solution from MILATRAS is inputted into SORTAH for solving of the user-constrained system optimal solution. The solution inputted from MILATRAS stochastically determines the passenger path sets, which models passenger travel behaviour. Though SORTAH deterministically solves for the user-constrained system optimal network, the stochastic input into the heuristic provides stochastic effects.

- Different standards of crowding levels are considered for each transit mode, because passengers have different tolerance levels for different transit modes. Passengers have better tolerance of crowding levels for vehicles that are designed for mostly standing passengers (i.e. subway) compared to vehicles designed for mostly seated passengers (i.e. buses). Therefore, the load factors ranges vary for each LOS rating (i.e. LOS A to LOS F).

### 5.2 Limitations

SORTAH has the following limitations:

- The procedure requires detailed information of each path in a passenger path set, including departure time, origin stop, destination stop, and all travel time components. Currently, it can only be applied to the user equilibrium solution from MILATRAS. However, future research can expand the theory to other user equilibrium solutions of microsimulations that contain detailed information of each passenger.

- The total travel time is an estimate and does not capture all network effects, because travel time in the model is given on passenger-by-passenger basis and summed up to yield the total travel time, while the network is analyzed on a segment-basis. Perceived travel time factor related to the load factor LOS cannot be applied because route segment
travel time cannot be extrapolated from the passenger information. When re-routing occurs, only the passenger who has been re-routed has their travel time updated. The impact on passengers that use the same segments are not captured due to the limitation of a model that does not calculate travel time on a segment-basis. Improvements on SORTAH to capture all network effects in the network’s total travel time can allow the hypothesis to be further examined.

- Temporal shifting within the hour is not modelled. For the same path, passengers are restricted to the specific departure times defined for the path. Although multiple transit vehicles can serve the path within the hour, passengers are modelled to depart and board the specific transit vehicle defined. For example, if a path that contains a route with 5-minute headways is modelled to serve the passenger at 7:15 am, the passenger would not be modelled and shifted to board the 7:10am or 7:20 am vehicle. Passengers are assumed to be on time so the transfers of the paths are retained.

- Departure time choice is ignored. It is assumed that passengers typically go from their origin to their destination within the same hour. They cannot be rerouted to another route of a different hour. For example, if the passenger is modelled to take a path between 6:00 am and 7:00 am, all 7 paths have departure times within the 6 AM hour-base.

- Transit vehicles are assumed to always be on-time, which does not capture true network effects. Transit networks typically have schedule adherence and on-time performance concerns that impact passenger travel experience. Transit vehicles that arrive at a stop/station outside of the scheduled time can cause passengers that are waiting for another transit vehicle to board the current vehicle and result in an influx of passengers bunching onto the same transit vehicle. As a result, this may cause a series of overcrowded segments.
6 Application Opportunities

6.1 Policies and Strategies

SORTAH can be used to test for policies and strategies that can achieve a constrained system optimum to reduce transit congestion. The traditional framework for evaluating policies and strategies is to first develop a policy or strategy and then test for its impact via a transit assignment model. The model outputs performance measures that compare the before and after scenarios against the goal of the policy. Depending on the outcome of the after-scenario, transit operators and planners would determine whether the policy or strategy would improve the current performance of the network. With this framework, transit operators and planners cannot determine whether their efforts are well justified when developing the policies and strategies. The motivation is to enhance the current state-of-the-art transit assignment model to determine the upper bound for which network performance can no longer improve and demand cannot be redistributed any further to relieve the congestion. Therefore, transit operators and planners can better concentrate their efforts on developing effective policies and strategies. This supports the proposed “reverse engineering” framework for evaluating policies and strategies whereby a network solution would drive the development of a policy or strategy.

6.2 ITS Applications

Emerging information, communication, sensor technologies and innovative transit operations control strategies are becoming critical elements of viable, competitive public transit systems. ITS technologies have significantly expanded the range of information available to the traveller through ATIS and improved the on-time performance of transit vehicles through the use of APTS. Passengers path choice can be influenced by ATIS, which can result in a user-constrained system optimal solution if deployed correctly. The SORTAH solution can be used as the reference and upper bound to examine technological effects on passenger route choice to develop an ATIS that drives towards a user-constrained system optimal solution. This supports the proposed “reverse engineering” framework for developing ITS technologies.
Currently, there are route guidance systems for auto mode that provide drivers with different path alternatives, typically assuming user equilibrium assumptions. There has been recent research that can upgrade route guidance systems that provide system-optimal paths for drivers (Jahn et al., 2005). Existing transit path finders can use SORTAH algorithm as a path of the route searching engine to influence passenger route choice towards a user-constrained system optimal solution. The transit mode route guidance system is a form of ATIS that can influence travel behaviour to drive the network towards a user-constrained system optimal solution.

6.3 Transit Route Design

Realistic passenger paths are bounded by the network design. Passengers typically consider between four to seven alternatives for travel choices and take paths with minimal transfers, minimal travel time and maximal comfort. Given the network design, passengers may be constrained to always travel through a section of a specific route for all path choices, because alternate routes may cause additional transfers, significant travel time, or discomfort from overcrowding for the passenger. The SORTAH solution can be used to determine and influence the physics design of new transit routes so the user equilibrium solution of the network would be equivalent to the user-constrained system optimal solution.
Chapter 5
Prototype Implementation

1 Chapter Overview
This chapter presents a proof-of-concept prototype implementation of SORTAH as an add-on feature of MILATRAS – Microsimulation Learning-based Approach to Transit Assignment.

- Section 2 provides a description of the prototypes.
- Section 3 describes the scope of analysis.
- Section 4 presents the results and discussion of the prototype implementation.
- Section 5 concludes with the most effective SORTAH procedure.

2 Prototype Description

2.1 MILATRAS
MILATRAS is an innovative transit assignment model developed by Wahba, which assigns passengers to the network according to their experience with the transit system performance (Wahba, 2009). Dynamic feedback of passenger’ trip choices and their adaptation to service performance are modelled using Markovian Decision Process (MDP) and Reinforcement Learning (RL). This microsimulation model assumes the user equilibrium principles. Prototype implementations of SORTAH were created as an add-on feature of MILATRAS to solve for the user-constrained system optimal solution by re-routing the passengers of the user equilibrium solution using the heuristic rules.
Figure 10 graphically presents the connection between MILATRAS and SORTAH. The prototype cannot be function as a stand-alone implementation and must be complemented with outputs from MILATRAS. First, the transit network, passenger’s initial choice sets, and expansion factors are developed to be inputted into MILATRAS. Using the mental model theory, MILATRAS solves for the user equilibrium solution, outputting passenger choices and their updated path set. The final passenger path set provides the user constraint element to the user-constrained system optimal solution. In addition, the original network representation and expansion factors are outputted by MILATRAS. Combining the outputs from MILATRAS with the LOS crowding standards and vehicle attributes developed for each transit mode, these become the input parameters that are fed into SORTAH. Following a set of procedures, SORTAH solves for the user-constrained system optimal solution.

**Figure 10. The SORTAH Prototype as an Add-On Feature of MILATRAS**

SORTAH performs all the heuristics steps following the re-routing criteria, the travel time adjustment on affected paths, and the re-routing evaluation. In total, there were three variations of the prototype developed to test the three different travel time criteria. All three prototypes still contain the three main sub-processes:

1. Pre-Process – Prepares the passenger path sets and route network data for re-routing.
2. Passenger Re-Routing Process – Re-routes passengers throughout the network using the re-routing criteria.
3. Number of Iterations – Determines the number of iterations of passenger re-routing process using the first type of re-routing evaluation. After the final iteration of the re-routing process, the second type of re-routing evaluation is applied.

However, the target travel time criterion in the passenger re-routing process is different. The purpose of these prototypes is to test for the most effective travel time criteria when solving for the user-constrained system optimal solution. The final SORTAH procedure would adopt the most effective algorithm for re-routing passengers that minimizes crowding levels and travel time after re-routing passengers.

2.2 Variations of the SORTAH Prototype

In conventional system optimal models, solutions are found by minimizing the total travel time which does not necessarily reduce crowding levels on the network. As discussed, the goal of SORTAH is to solve for a network in which there are no more “pinch points” or overcrowded routes in the network. This thesis hypothesizes that total travel time will indirectly be minimized when crowding levels are directly minimized.

Three variations of the SORTAH prototype were developed to test the hypothesis. Each variation of the SORTAH prototype differs by the target travel time criterion implemented. They range from the strictest to relaxed travel time criterion:

- **Case 1 – Shorter Time Check:** Target path has smaller travel time than original path.
- **Case 2 – Measure of Unfairness:** Target path has smallest measure of unfairness.
- **Case 3 – No Time Check:** Travel time criterion is relaxed.

For each test case, the travel time criterion is applied after it has been determined that re-routing to another path is possible.

Figure 11 presents the SORTAH procedure using the strictest travel time criterion that guarantees a reduction in the total travel time. Passengers are re-routed to a path with a travel
time that is smaller than the original. In this case, crowding levels and travel time factors are directly minimized.

**Figure 11. Case 1 – Shorter Time Check**

Figure 12 illustrates the SORTEA procedure using a modified travel time criterion that does not guarantee a reduction in the total travel time. Passengers are re-routed to a path with the smallest measures of unfairness. Though re-routing passengers to a path with the smallest measures of unfairness does not guarantee a reduction in the total travel time, it minimizes the marginal total travel time. The marginal total travel time ranges from negative to positive. Minimizing the marginal total travel time would drive the value as negative as possible to reduce the total travel time. However, if the marginal travel time is positive, the total travel time increase is minimized. In this case, crowding levels are directly minimized and the total travel time is indirectly
minimized. However, there is added procedure to guide the total travel time for reduction, even though it is not guaranteed.

**Figure 12. Case 2 – Measure of Unfairness Check**

Figure 13 demonstrates the SORTAH procedure with the travel time criterion relaxed. Passengers are only re-routed to a path that meets the target crowding level criterion. All travel time related factors are ignored in the re-routing process. This case tests the hypothesis of whether directly minimizing crowding levels will indirectly minimize the total travel time. Comparing the results of Case 3 with the previous two cases would show the effects of the shorter travel time criterion, measure of unfairness criterion and no travel time criterion on the user-constrained system optimal solution.
Figure 11, Figure 12, and Figure 13 illustrate the difference between the three variations of the SORTAH prototype. The varying travel time criteria relative to rest of the SORTAH procedure are highlighted.

Figure 13. Case 3 – No Travel Time Check

- Calculate original total network travel time
- Calculate load factors of each segment.
- Assign LOS rating to each segment.
- Create a list of LOS F segments.
- Update LOS F segments list.
- Re-route passenger to target path.
- Adjust travel time for target path (-2s).
- Will these paths have LOS F segments when he/she is re-routed to this path?
- Are there other paths in his/her set without LOS F segments?
- Create a list of passengers who uses a path with LOS F segment(s).
- Choose a passenger.
- Calculate new number of LOS F segments.
- Calculate new total network travel time.
- Are there more passengers in the passenger list?
- Create a list of passengers who uses a path with LOS F segment(s).
- Choose a passenger.
- Calculate new number of LOS F segments.
- Calculate new total network travel time.
- Are there more passengers in the passenger list?
- Re-route passenger to target path.
- Adjust travel time for the original path (+2s).
- Adjust travel time for the target path (-2s).
3  Scope of Analysis

Three variations of the SORTAH prototype were developed to test the hypothesis that total travel time will indirectly be minimized when crowding levels are directly minimized. The hypothesis is tested by comparing each case against the elements of analysis below:

- Crowding levels – Measured by the number of LOS F segments.
- Total travel time – Measured by the network total travel time.
- Execution time – Measured by the time it takes to run the prototype, relative to the network size.

All modelling and analysis are carried out assuming the morning peak period, which spans from 6:00 am to 9:00 am. Crowding levels for each route and segment are measured to evaluate the effectiveness of each heuristic in reducing overcrowding on the network. Total travel time is measured to examine the effectiveness of each travel time criterion in decreasing the total network travel time. The combined results of the measures of crowding and perceived travel time measures provide an overview of how operational efficiency has improved. The desired heuristic of SORTAH is one in which crowding levels and total travel time has reduced the most.

The purpose of the heuristic is to assist planners in making a quick and well-informed decision on policy implementation. The execution time measure evaluates how fast each case can solve for the user-constrained system optimal solution. Real-world application of SORTAH requires the heuristic to find a good solution quickly. The desired heuristic of SORTAH is one in which execution time is no more than a few hours.
4 Results and Discussion

This section presents the results and discussion of the analyses conducted for each variation of the SORTAH prototype developed. Analyses were carried out for the 3-hour morning peak period from 6:00 am to 9:00 am for the TTC network. Each case was compared for their effectiveness in reducing crowding levels and total travel time. Execution time was also evaluated to measure the program’s performance.

4.1 Crowding Levels

Measures of overcrowding examine the operational efficiency of a network with respect to crowding and comfort level of a network on a segment- and route-basis. Three measures of overcrowding were examined: measure of route overcrowding, measure of route overcrowding change and measure of segment overcrowding change.

Measure of route overcrowding assesses the number of routes affected by congestion. In other words, congestion is quantified by the number of routes with LOS F segments. Table 12 compares the number of routes with LOS F segments between each prototype during the 3-hour morning peak period. Each affected route is further categorized for whether they have been congested for one hour, two hours, or three hours during the peak period.

Table 12. Measure of Route Overcrowding

<table>
<thead>
<tr>
<th># of Hours</th>
<th># of Overcrowded Routes</th>
<th>$\Omega_{\text{route}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Base</td>
<td>Case 1</td>
</tr>
<tr>
<td>1 hour</td>
<td>16</td>
<td>15</td>
</tr>
<tr>
<td>2 hours</td>
<td>14</td>
<td>19</td>
</tr>
<tr>
<td>3 hours</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>Overall</td>
<td>34</td>
<td>35</td>
</tr>
</tbody>
</table>

Total Number of Routes: 148

The base case shows the network congestion of the user equilibrium solution. There were 16 routes with LOS F segments for one hour, 14 routes with LOS F segments for two hours, and four routes with LOS F segments for three hours. In total, there were 34 overcrowded routes.
These numbers are used to determine the measure of route overcrowding by dividing the number of overcrowded routes by the total number of routes in the system. Overall, 23% of the routes in the network were overcrowded. Case 1 is the prototype that uses the travel time criterion where the target path has smaller travel time than the original path. There were 15 routes with LOS F segments for one hour, 19 routes with LOS F segments for two hours, and one route with LOS F segments for three hours. Re-routing of passenger using Case 1 criterion reduced the number of overcrowded routes for the one-hour and three-hour duration, but increased the number of overcrowded routes for the two-hour duration. The overall effect increased the total number of overcrowded routes by one to 35 overcrowded routes. Overall, 24% of the routes in the network were overcrowded, which is slightly higher than the base case. Case 2 is the prototype that uses the travel time criterion where the target path has smallest measure of unfairness. There were 12 routes with LOS F segments for one hour, 12 routes with LOS F segments for two hours, and no routes with LOS F segments for three hours. For all three durations, the number of overcrowded routes decreased. In total, there were 24 overcrowded routes. Overall, 16% of the routes in the network were overcrowded, which reduced from the user equilibrium network. Case 3 is the prototype that does not use any travel time criterion. It yielded the same results as Case 2 in which there were 12 routes with LOS F segments for one hour, 12 routes with LOS F segments for two hours, and no routes with LOS F segments for three hours.

By re-routing passengers to a target path with the smallest measure of unfairness or not using any travel time criterion yields similar results. Both cases reduced the number of overcrowded routes by 29.41% from the user equilibrium network. On the other hand, re-routing passengers to a target path that has a smaller travel time than original path increased the number of overcrowded routes by 2.94%. These results indicate that Case 2 and Case 3 are more effective in improving the operational efficiency of the network with respect to crowding and comfort level of a network on a route-basis. Table 13 outlines the percentage change in route overcrowding from the base case.
Table 13. Percentage Change in Route Overcrowding

<table>
<thead>
<tr>
<th># of Hours</th>
<th>Case 1</th>
<th>Case 2</th>
<th>Case 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 hour</td>
<td>-6.25%</td>
<td>-25.00%</td>
<td>-25.00%</td>
</tr>
<tr>
<td>2 hours</td>
<td>+35.71%</td>
<td>-14.29%</td>
<td>-14.29%</td>
</tr>
<tr>
<td>3 hours</td>
<td>-75.00%</td>
<td>-100.00%</td>
<td>-100.00%</td>
</tr>
<tr>
<td>Overall</td>
<td>+2.94%</td>
<td>-29.41%</td>
<td>-29.41%</td>
</tr>
</tbody>
</table>

Measuring operational efficiency of the network with respect to crowding and comfort level of a network on a route-basis provides a high-level indication of how well each variation of prototype can reduce overcrowding on the network. Thought it indicates the intensity of overcrowding with respect to the duration of congestion, it does not provide the number of pinch points without the route that is cause it to be overcrowded. For example, a route with 10 segments could have four overcrowded segments for a period of one hour, while another route with 12 segments could have nine segments for a period of one hour. The latter route would be more congested because it has more “pinch points” than the former route. Roughly 40% of the former route is congested compared to 75% of the latter route. By measuring congestion on a segment-basis, granularity and intensity of network overcrowding can be quantified. Figure 14 illustrates the number of LOS F segments in the network for each iteration and Table 14 shows the percentage change in segment overcrowding from the user equilibrium network.

As shown in the graph below, the base case had roughly 1700 LOS F segments. By re-routing passengers to a target path with shorter travel time (i.e. Case 1), the number of LOS F segments reduced by 33.53% in 13 iterations. In Case 2, passengers were re-routed to a path with the smallest measure of unfairness. The number of LOS F segments decreased by 59.28% in 18 iterations. In Case 3, the travel time criterion is relaxed and reduced the number of LOS F segments by 59.28% in 9 iterations. Case 2 and Case 3 reduces the number of LOS F segments two times more than Case 1. This indicates that Case 2 and Case 3 are more effective in improving the operational efficiency of the network with respect to crowding and comfort level of a network on a segment-basis.
Figure 14. Comparing Crowding Levels for each Case

Table 14. Percentage Change in Segment Overcrowding

<table>
<thead>
<tr>
<th></th>
<th>Case 1</th>
<th>Case 2</th>
<th>Case 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\Delta_{\text{route}}$</td>
<td>-33.53%</td>
<td>-59.28%</td>
<td>-57.62%</td>
</tr>
</tbody>
</table>

When passengers are re-routed to a target path with a shorter travel time, the number of overcrowded routes increased while the number of LOS F segments decreased. This indicates that the overall network congestion effects decreased and the average overcrowding of each route has also reduced. However, the number of overcrowded routes is still the same and all the routes have not been eliminated of overcrowding. On the other hand, when passengers are re-routed to a target path with the smallest measure of unfairness or no travel time criterion, the number of overcrowded routes and the number of LOS F segments decreased. The overall network congestion effects decreased and several of the overcrowded routes were eliminated.
The average overcrowding of each route is also reduced. Case 2 and Case 3 heuristic reduces crowding and comfort level of a network two times better than Case 1 on route- and segment-basis.

4.2 Total Travel Time

Total travel time is used to assess the operational efficiency of a network with respect to travel time. The total travel time change is only applicable because demand and supply is constant, therefore user equilibrium solution can be compared to the user-constrained system optimal solution of the transit network. Otherwise, normalized measure of the total travel time should to be used for the same metric base to compare two networks of different supply and demand. In this case, both total travel time and average passenger travel time are reported for better comparison of each variation of the prototype to improve operational efficiency with respect to travel time. Figure 15 illustrates the total travel time of the network for each iteration and Table 15 shows the percentage change in total travel time from the user equilibrium network.

The user equilibrium network had a total travel time of roughly 99.6 thousands hours (or 5,977,073 minutes). On average, each passenger took 18 minutes to travel from their origin to their destination. In Case 1, passengers were re-routed to a target path with shorter travel time, which reduced the total travel time by 0.64% to 99.0 thousand hours (or 5,938,616 minutes) in 13 iterations. This is equivalent of saving each passenger 0.1 minute (or 6 seconds) for each trip. In Case 2, passengers were re-routed to a path with the smallest measure of unfairness. This increased the total travel time by 1.12% to 100.7 hours (or 6,044,133 minutes) in 9 iterations. Passengers would be taking an additional 0.2 minute (or 12 seconds) for each trip. In Case 3, the travel time criterion was relaxed, thereby increasing the total travel time of the network 2.24% to 101.9 thousand hours (or 6,111,202 minutes) in 18 iterations. In other words, passengers would be taking an additional 0.5 minute (or 30 seconds) for each trip.
Figure 15. Comparing Total Travel Time for each Case

Table 15. Percentage Change in Total Travel Time

<table>
<thead>
<tr>
<th></th>
<th>Base</th>
<th>Case 1</th>
<th>Case 2</th>
<th>Case 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total Travel Time (min)</td>
<td>5,977,073</td>
<td>5,938,616</td>
<td>6,044,133</td>
<td>6,111,202</td>
</tr>
<tr>
<td>Average Passenger Travel Time (min)</td>
<td>18.0</td>
<td>17.9</td>
<td>18.2</td>
<td>18.5</td>
</tr>
<tr>
<td>$\Delta_t$</td>
<td>-</td>
<td>-0.64%</td>
<td>+1.12%</td>
<td>+2.24%</td>
</tr>
</tbody>
</table>

The changes in the average passenger travel time indicate that these re-routing strategies have little impact on changes in the total travel time and passenger travel experience. Though Case 1 uses a stricter travel time criteria and reduces the total travel time, the 6 seconds in which a passenger would save for their trip is very miniscule and have little impact on their travel experience. Similarly, the travel time increase of 12 seconds and 30 seconds per trip for Case 2 and Case 3, respectively, is too small to have significant impact on passenger travel experience.
Passengers would start noticing changes in their travel time when the marginal travel time is two minutes or more. Though Case 2 and Case 3 increase the travel time, Case 2 is preferred because the use of the measure of unfairness as the travel time criterion to re-route passengers minimizes the marginal total travel time. On the other hand, Case 3 does not directly control travel time when the travel time criterion is relaxed.

These results do not support the hypothesis in which directly minimizing crowding levels would indirectly minimize travel time. However, this does not disprove the hypothesis. Since SORTAH currently analyzes the network on a segment basis, the total travel time is an estimate and does not capture all network effects. Travel time in the model is given on passenger-by-passenger basis and summed up to yield the total travel time. When re-routing occurs, only the passenger who has been re-routed has their travel time updated. Impact on passengers that use the same segments are not captured due to the limitation of the model that does not calculate travel time on segment-basis. Improvements on SORTAH to capture all network effects in the network’s total travel time can allow the hypothesis to be further examined.

4.3 Execution Time

Execution time measures how fast each variation of the prototype can solve for the user-constrained system optimal solution. In addition to improving the operational efficiency of a network with respect to crowding levels and travel time, the purpose of the heuristic is to assist planners in making a quick and well-informed decision on policy implementation.

Figure 16 compares the execution time for each case relative to the network size. For all three variations, the execution time increases exponentially when network size increases between 0 to 50 branches and then plateaus when network size is greater than 50 branches. Network size is measured by the number of branches rather than the number of routes because each direction and branch of a route needs to be considered when SORTAH solves for the user-constrained system optimal solution. Re-routing of passengers occur by the branches of routes. For example, passengers would not be re-routed from a route going northbound to the same route going southbound.
When passengers were re-routed to a target path with shorter travel time (i.e. Case 1), the execution time for a network of roughly 480 branches was 77 minutes. In Case 2 where passengers were re-routed to a path with the smallest measure of unfairness, the execution time was 85 minutes. For Case 3 where no travel time criterion was used, the execution time was 45 minutes. When the travel time criterion is relaxed when re-routing passengers, it reduced the execution time by half, as shown in the graph. Case 2 and Case 3 execution times were comparable, but the execution time for Case 2 is slightly slower than Case 3. Though Case 1 runs within one hour, it does not improve operational efficiency with respect to travel time and crowding levels as well as Case 2 and Case 3.
5 Conclusion

The user equilibrium solution served as the base case to compare three variations of the travel time criteria: Case 1 (Shorter Time Check), Case 2 (Measure of Unfairness), and Case 3 (No Time Check). In the base case, 34 out of 148 routes were overcrowded and there were approximately 1700 overcrowded segments. The total network travel time was 99.5 thousand hours, which equates to a passenger taking roughly 18 minutes to complete his/her trip. For each variation of prototype, the crowding levels and total travel time changed from the base case accordingly. Table 16 summarizes the percentage change in crowding levels and total travel time for each case.

Table 16. Summary of Percentage Change in Crowding Levels and Total Travel Time

<table>
<thead>
<tr>
<th></th>
<th>Case 1</th>
<th>Case 2</th>
<th>Case 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Δ_route</td>
<td>+2.94%</td>
<td>-29.41%</td>
<td>-29.41%</td>
</tr>
<tr>
<td>Δ_segment</td>
<td>-33.53%</td>
<td>-59.28%</td>
<td>-57.62%</td>
</tr>
<tr>
<td>Δ_tt</td>
<td>-0.64%</td>
<td>+1.12%</td>
<td>+2.24%</td>
</tr>
</tbody>
</table>

The results indicate that the use of shorter time check as the travel time criterion does not perform as well as the use of measure of unfairness or no time check. When passengers are re-routed to a path with a shorter travel time, the number of overcrowded routes increased while the number of overcrowded segments decreased. This is an indication that Case 1 (Shorter Time Check) spreads out overcrowding amongst the network routes rather than eliminating as many overcrowded routes as possible. When passengers are re-route to a path with the smallest measure of unfairness, both the number of overcrowded routes and number of overcrowded segments decreased. The overall congestion on the transit network is decreased by eliminating as many overcrowded routes as possible. Similarly, when travel time criterion is relaxed during the re-routing process, the number of overcrowded routes and number of overcrowded segments decreased. Case 2 (Measure of Unfairness) and Case 3 (No Time Check) improves operational efficiency with respect to comfort and crowding levels. However, the use of the measure of unfairness reduces transit congestion slightly better than no use of travel time criterion.
When passengers are re-routed using either re-routing strategies, the change in average passenger travel time and total travel time have little impact on the passenger travel experience. Case 1 (Shorter Time Check) that uses a stricter travel time criteria saves each passenger an average of 6 seconds per trip. Case 2 (Measure of Unfairness) and Case 3 (No Time Check) increase an average of 12 seconds and 30 seconds, respectively, per trip. These changes in travel time have little impact on passenger’s travel experience because they are too small for passengers to notice a difference. Passengers would start noticing changes in their travel time when the marginal travel time is two minutes or more. Though Case 2 (Measure of Unfairness) and Case 3 (No Time Check) increase the travel time, the use of the measure of unfairness as the travel time criterion is preferred because it minimizes the marginal total travel time. When travel time criterion is relaxed, fluctuations in travel time is not directly controlled like in Case 1 (Shorter Time Check) or Case 2 (Measure of Unfairness).

Though, these results do not support the hypothesis in which directly minimizing crowding levels would indirectly minimize travel time, it does not disprove the hypothesis. SORTAH currently analyzes the network on a segment basis; therefore, the total travel time is an estimate and does not capture all network effects. Travel time in the model is given on passenger-by-passenger basis and summed up to yield the total travel time. When re-routing occurs, only the passenger who has been re-routed has their travel time updated. Impact on passengers that use the same segments are not captured due to the limitation of the model that does not calculate travel time on segment-basis. When improvements on SORTAH are available to capture all network effects in the network’s total travel time, the hypothesis can be further examined.

Real-world application of SORTAH requires the heuristic to find a good solution to assist planners in making a quick and well-informed decision on policy implementation. The execution time measure evaluates how fast each variation of the prototype can solve for the user-constrained system optimal solution. Results show that Case 3 (No Time Check) had the fastest execution time of less than one hour. This is because the travel time criterion is relaxed and there are fewer criteria to check during the re-routing process. Case 1 (Shorter Time Check) or Case 2 (Measure of Unfairness) have comparable execution times which are approximately two times
the execution of Case 3 (No Time Check). All three cases have desirable execution time because they do not exceed more than two hours. Table 17 summarizes and compares the results of the three variations of the prototype.

The final SORTAH procedure adopts the algorithm of Case 2 (Measure of Unfairness). Passengers are re-routed to paths with the smallest measure of unfairness. It has been found that this heuristic procedure provides better operational efficiency with respect to comfort, crowding levels and travel time. Its execution time is no more than two hours and quickly solves for the user-constrained system optimal solution that minimizes crowding levels and travel time.
<table>
<thead>
<tr>
<th>Case #</th>
<th>Travel Time Criterion</th>
<th>Case 1</th>
<th>Case 2</th>
<th>Case 3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Shorter Time Check</td>
<td>Measure of Unfairness</td>
<td>No Time Check</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Congestion increased on a route-basis</td>
<td>Congestion decreased on a route-basis</td>
<td>Congestion decreased on a route-basis</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Overall congestion on segment-basis decreased, improving passenger travel experience</td>
<td>Overall congestion on segment-basis decreased, significantly improving passenger travel experience (e.g. decrease number of overcrowded segments by 59.28%)</td>
<td>Overall congestion on segment-basis decreased, significantly improving passenger travel experience (e.g. decrease number of overcrowded segments by 57.62%)</td>
</tr>
<tr>
<td></td>
<td><strong>Comparison</strong></td>
<td>Travel time on reduces slightly, but the change is very miniscule (e.g. 6-second time savings per trip)</td>
<td>Travel time increased slightly, but the change is very small to have significant impact on passenger’s travel experience (e.g. 12-second time increase per trip)</td>
<td>Travel time increased, but the change may or may not have impact on passenger’s travel experience (e.g. 30-second time increase per trip)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Minimizes the total travel time increase</td>
<td>Does not control the total travel time increase, because there are not travel time criterion for re-routing</td>
<td>Quick execution time of less than one hour even for large-scale networks with 480 branches (e.g. 45 minutes)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Execution time that exceeds one hour for large-scale networks with 480 branches, but still quick in solving for the user-constrained system optimal solution (e.g. 77 minutes)</td>
<td>Execution time that exceeds one hour for large-scale networks with 480 branches, but still quick in solving for the user-constrained system optimal solution (e.g. 85 minutes)</td>
<td>Comparable to Case 2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Comparable to Case 2</td>
<td>Comparable to Case 1</td>
<td>Comparable to Case 1</td>
</tr>
</tbody>
</table>
Chapter 6
Case Study – The Toronto Transit Commission (TTC) Application

1 Chapter Overview

This chapter presents a large-scale real-world application of SORTAH on the Toronto Transit Commission (TTC) network.

- Section 2 describes the application background.
- Section 3 presents the data requirements.
- Section 4 outlines the scope of analysis.
- Section 5 discusses and analyzes the results of SORTAH on the TTC network.
- Section 6 concludes with the findings.

2 Application Background

2.1 The Transportation Tomorrow Survey (TTS)

Since 1986, comprehensive travel surveys have been conducted in the Greater Toronto Area (GTA) and surrounding areas every five years (i.e. 1991, 1996, 2001, 2006). This survey is known as the Transportation Tomorrow Survey (TTS), which collects detailed demographic and travel information on all household members for 5% of all households in the GTA and surrounding regions shown in Figure 17. Each region is identified with a range of zone codes detailed in Table 18. For example zone codes in the City of Toronto range from 1 to 481. The purpose is to use this data to understand travel behavior in the GTA, which is an area divided into five regions: City of Toronto, Durham, Peel, York, and Halton.
Figure 17. GTA and Surrounding Regions for Year 2001 (Joint Program in Transportation, 2003a)

Table 18. GTA and Surrounding Regions’ Zone-Coding for Year 2001 (Joint Program in Transportation, 2003a)

<table>
<thead>
<tr>
<th>Zone-Code</th>
<th>Area</th>
<th>Zone-Code</th>
<th>Area</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 – 481</td>
<td>Toronto</td>
<td>3601 – 3650</td>
<td>Guelph</td>
</tr>
<tr>
<td>501 – 765</td>
<td>Durham</td>
<td>3707 – 3716</td>
<td>Wellington</td>
</tr>
<tr>
<td>1001 – 1353</td>
<td>York</td>
<td>3717 – 3720</td>
<td>Dufferin</td>
</tr>
<tr>
<td>1501 – 1753</td>
<td>Peel</td>
<td>3721 – 3726</td>
<td>Orangeville</td>
</tr>
<tr>
<td>2001 – 2197</td>
<td>Halton</td>
<td>3731 – 3735</td>
<td>Orillia</td>
</tr>
<tr>
<td>2501 – 2670</td>
<td>Hamilton</td>
<td>3741 – 3793</td>
<td>Simcoe</td>
</tr>
<tr>
<td>3001 – 3087</td>
<td>Niagara</td>
<td>3801 – 3832</td>
<td>Barrie</td>
</tr>
<tr>
<td>3201 – 3499</td>
<td>Waterloo</td>
<td>3841 – 3857</td>
<td>Kawartha Lakes</td>
</tr>
<tr>
<td>3501 – 3549</td>
<td>Brantford</td>
<td>3871 – 3875</td>
<td>Peterborough</td>
</tr>
<tr>
<td>3550 – 3560</td>
<td>Brant</td>
<td>3901 – 3925</td>
<td>Peterborough City</td>
</tr>
</tbody>
</table>
The data used in this application refer to the collected TTS records in the year 2001 (TTS2001) and specifically transit trips served by the Toronto Transit Commission (TTC).

The TTC is a public transport agency that operates transit bus, streetcar, paratransit, and rapid transit services within the City of Toronto. Toronto is considered the economic heart of the GTA. In 2001, it was reported that 45% of the GTA work trips are destined to Toronto alone, with a 25% modal split for transit (Joint Program in Transportation, 2003a). Data collected includes travel information about the trips made by the household members over an entire weekday and transit trip data that includes trip start time, trip purpose, origin and destination geo-locations, and the sequence of transit routes for those trips. While the survey covers only 5% of all the households in the GTA, expansion factors are used to expand the information to cover all of them. The expansion factors are determined and verified based on Census Sub-Division level data (Joint Program in Transportation, 2003b). In 2001, 85,095 individuals across the GTA reported that they used transit as their mode of travel, which expanded to 1,469,237 transit trips across the GTA for a weekday. Out of the total number of transit trips, it has been reported that 78% of those trips use the TTC service and over 332,000 trips occurred in the morning peak period. Given these statistics, the TTC system was selected for the development of the large-scale model due to its importance and significance to the GTA transportation network.

### 2.2 The Toronto Transit Commission (TTC) System

The TTC network is a transit system of surface routes (such as buses and streetcars) and underground subway lines that serves the City of Toronto. It operates four different types of services:

1. Traditional bus service (e.g. Route # 85 – Sheppard West)
2. Express bus service (e.g. Route # 141 – Downtown Express Route)
3. Light Rail Transit/Street service (e.g. Route # 510 – Spadina Streetcar)
4. Rapid Rail Transit/Subway service (e.g. Route # 601 – Bloor-Danforth Line)
In 2001, the TTC operated 294 one-directional routes (or 147 two-directional routes) during the morning/AM peak period (TTC, 2001). Since some of these routes run on different branches, covering different segments; therefore a total of 480 branches were modelled. During the AM period, these branches serve roughly 10,000 stops. The frequency of service over these branches spans from high frequency service (2-minute headway) to low frequency service (60-minute headway); medium frequency services are represented by different values.

Though the TTC operates within the Toronto jurisdictions, some interactions with neighboring transit systems exist along the city boundary. Transfers between Mississauga Transit or York Region Transit systems and the TTC service are common. It has been recorded by TTS that about 45% of all work trips for all modes of travel destined to Toronto and 31% of those trips originated in the GTA regions surrounding Toronto (i.e. Peel, York, and Durham). Roughly 12% of transit trips that use TTC in the AM peak period originated outside the Toronto boundaries and about 4% of the transit trips that use TTC in the AM peak period were destined for outside the Toronto boundaries. The modelling boundary of the case study is not restricted by the physical boundary of the TTC service. The application deals with trips originating outside the Toronto boundary at their origin and trips destined for outside the Toronto boundary at their destination.

The TTC system interacts with inter-regional rail and bus services in the GTA, represented by the Government of Ontario (GO) service. The GO network covers large parts of the GTA, which further expands the geographical boundary of the case study. During the peak period, GO operates a number of heavily used inter-regional train and bus routes. There are major TTC/GO connection points within the Toronto area that allow for a great deal of interaction. For example, Union Station is a major hub for the TTC service and also a major GO hub, which is located in the downtown Toronto area. Since GO services are not frequent and are largely dependent on the time of day and area of coverage, there are temporal-constraints on the possible connections that GO users may consider to transfer to and from TTC service. The usage of the TTC service by GO riders is time-dependent and is modelled dynamically. Also, the different fare structures that TTC and GO operate with, are modelled to capture these interactions.
3 Scope of Analysis

The scope of analysis for this case study is as follows:

- All data used in the application reflect the year 2001 conditions, unless stated otherwise. The current TTC conditions are not reflected, but the purpose of the analysis is to provide a proof-of-concept, which can be expanded to other large-scale networks.

- All modelling and analyses are carried out for the TTC network during the AM peak period only. The AM peak period in this case study spans from 6:00 am to 9:00 am.

- A trip-end is considered within the TTC service coverage area if it is located within the City of Toronto boundary, or it is located within the maximum walking distance from a TTC service stop. GO riders who originate outside the City of Toronto are modelled to feed into the TTC system.

- SORTAH that adopts the algorithm of Case 2 (Measure of Unfairness) identifies the overcrowded routes before and after the re-routing process. Further analysis is conducted to determine whether an increase in capacity is required for specific TTC routes after re-routing has been attempted.

4 Data Requirements

There are two main data sources fed into SORTAH for modelling: MILATRAS outputs and re-routing criteria. The outputs of MILATRAS provide each passenger travel experience and the details of the supply network. SORTAH reorganises these outputs into an analysable form for SORTAH to conduct the re-routing process. The re-routing criteria provide SORTAH with standards for when re-routing should occur.
4.1 MILATRAS Outputs

MILATRAS uses three data sources to model the demand and supply of the TTC network: Data Management Group (DMG) at the University of Toronto, TTC Service Planning Department, and the University of Toronto Library. Each source provided different types of data to be used for various stages in the MILATRAS modelling process, thereby outputting passengers’ travel experience and details of the network for SORTAH.

4.1.1 Demand

The DMG at the University of Toronto is the custodian of the data sets collected and derived from the TTS. It provides online access to the TTS data sets (i.e. derived data and not raw data collected records) through the Data Retrieval System (iDRS) using individual accounts that are granted based on request. By applying a set of filters to TTS2001 data set, all the trips that use the TTC and have a start time variable between 6:00 am and 9:00 am were tabulated by their origin and destination zone information. The output of this procedure is an origin-destination matrix (OD matrix) of 332,073 trips that can be further divided by the access mode of the trip to the transit service and the egress mode to the destination from the transit service. The modelled access and egress modes are ‘walk’, ‘passenger’, or ‘driver’. The ‘walk’ mode would have a smaller access or egress speed than the ‘passenger’ or ‘driver’ mode. These are accounted for in the access and egress times of each individual’s travel time. Table 19 outlines the access and egress modes frequency and joint distributions for the AM peak period in 2001.

<table>
<thead>
<tr>
<th>Access Modes</th>
<th>Egress Modes</th>
<th>Total Access</th>
</tr>
</thead>
<tbody>
<tr>
<td>Walk</td>
<td>87.89%</td>
<td>0.06%</td>
</tr>
<tr>
<td>Passenger</td>
<td>5.00%</td>
<td>0.01%</td>
</tr>
<tr>
<td>Driver</td>
<td>6.60%</td>
<td>0.06%</td>
</tr>
<tr>
<td>Total Egress</td>
<td>99.49%</td>
<td>0.45%</td>
</tr>
</tbody>
</table>

Table 19. Access and Egress Modes Frequency and Joint Distributions for 332,073 Trips
MILATRAS also models the OD matrix by trip purpose. In the TTS2001 data set, the AM peak period consisted of:

- 67% home-based work trips
- 27% home-based school trips
- 4% home-based other trips
- 2% non-home-based trips

These trips were determined by applying the land use maps for the Province of Ontario onto the OD matrix layer. The land use maps were obtained from the University of Toronto Map Library. Potential destination locations were categorised as commercial, resources and industrial, or government and institutional. Potential origin locations were marked as residential.

MILATRAS does not restrict the modelling boundary of the case study by the physical boundary of the TTC service. Passengers who are originated outside the Toronto boundary at their origin or are destined for outside the Toronto boundary at their destination interact with inter-regional rail and bus services in the GTA (GO network) and neighboring transit systems that exist along the city boundaries. MILATRAS dynamically models the usage of the TTC service by GO riders. SORTAH takes into account these GO riders who also use the TTC for the morning peak commute. However, only the TTC travel portion of these passengers is modelled in SORTAH and accounted for when measuring overcrowding on the segments and routes.

These characteristics of demand for TTC allow MILATRAS to simulate passenger path choices that reflect the real-world network effects. SORTAH takes the mental model output of MILATRAS as the demand input. These mental model data structures contain each passenger’s travel experience, in which SORTAH reorganizes into a comma-separated value file format for easier analysis and re-routing purposes.
4.1.2 Supply

MILATRAS uses the data provided by the TTC Service Planning Department and the University of Toronto Map Library to form the TTC and GO network for supply modelling. The data used for the TTC network reflect the 2001 conditions and the GO network reflect the 2002 conditions. The 2001 conditions of the GO network could not be found; therefore the 2002 conditions were used. Details of each route included the route name and number, number of branches, the coverage area, service timetable, a complete list of route-stop sequences, and Geographical Information System (GIS) shape files for the routes and stops. This data was reorganized and compiled together, to represent the network as a linked graph with elements of stops, route segments, routes, bounds, branches, vehicle types and transfer group. SORTAH uses this supply network to discretize each passenger’s paths into segments for disaggregated analysis. The crowding levels can be measured and “pinch points” of the network can be identified.

4.2 Re-Routing Criteria

There are four re-routing criteria that need to be met for passengers to use an alternate path. Of the four criteria, the current crowding level criterion and target crowding level criterion require standardized load factor ranges for different LOS grades for re-routing purposes. These standards are transit mode dependent, meaning the load factor ranges for different LOS grades vary depending on whether the passenger is travelling by bus, subway, or streetcar. These standards are represented by the load factor file that is inputted into SORTAH. Program users of SORTAH have the flexibility to change, add, or remove the standards for different transit modes as required. SORTAH also provides that flexibility for program users to customise the definition of overcrowding by transit mode. Currently, overcrowding is defined by any segment or route with LOS F segments. Passengers are re-routed away from these segments with LOS F. However, program users of SORTAH can choose to tighten the definition of overcrowding by re-routing passengers away from segments with LOS E or above by transit mode.
5 Results and Discussion

The analysis of the TTC network is based on the TTS2001 AM peak period data. It serves as a proof-of-concept to examine the applicability of SORTAH on large-scale networks. The results reflect the output of SORTAH that adopts the algorithm of Case 2 (Measure of Unfairness). It has been found that none of the subway routes have reached LOS F as we observe today and the TTC network still experiences overcrowding after the re-routing process. Figure 18 summarizes the number of overcrowded routes before and after the re-routing process.

**Figure 18. Number of Overcrowded Routes Before and After Re-Routing Process**

This graph identifies which the overcrowded routes or critical routes that TTC should focus their strategy on. It summarizes the network on a route-by-route basis. The “before re-routing” scenario illustrates the number of overcrowded routes in a user equilibrium solution and the “after re-routing” scenario shows the number of overcrowded routes in a user-constrained system optimal solution. In the user equilibrium solution, there were 34 overcrowded routes out of the
148 routes that serve the TTC network. All these overcrowded routes were bus routes, and none of the subway lines or streetcar routes was experiencing overcrowding in 2001. This is contrary to today’s observations of the TTC system because reports are showing congestion on the subway lines and streetcar routes, especially during the peak periods. The five bus routes that experienced the most overcrowding in the user equilibrium solution were Routes 35, 45, 54, 85, and 102. After applying the SORTAH algorithm, 10 out of the 34 overcrowded routes were relieved of congestion. Congestion was completely eliminated for TTC Bus Routes 11, 12, 60, 61, 63, 68, 70, 79, 96, and 98. The re-routing process also reduced overcrowding on Routes 35, 45, and 102 by more than half. However, Routes 54, and 85 remained heavily congested. This could be an indication that there are no existing alternate routes in which passengers can be re-routed to. Even with the deployment of ITS technologies that can influence passenger path choice towards a user-constrained system optimal solution to improve operational efficiency, passengers cannot be re-routed to an alternate path. Capacity would need to be added on to the heavily congested routes, or a relief route would need to be added to the existing network to accommodate the demand and eliminate overcrowding on the critical routes.

Further analysis was conducted on segment-by-segment basis for each critical route identified. From these heavily congested routes, “pinch points” on a network were identified, which can assist the TTC agency to develop a policy or strategies that can relieve congestion that these critical points. These disaggregated analyses were conducted using a passenger load profile diagram. These diagrams are specific to the route, route direction, transit mode, and the discretized analysis period (i.e. one-hour time period base, such as the period between 7:00 am and 8:00 am within the AM peak period). An example of this disaggregated analysis is shown in Figure 19 and Figure 20.

Figure 19 shows a passenger load profile diagram for Southbound Route 102 from 7:00 am to 8:00am. Route 102 is a bus route and therefore the LOS grades are highlighted with the corresponding load factor ranges. In the “before re-routing” scenario, it has been found passengers gradually board the train from stop 5586 to stop 5559 (from the right end to the middle of the graph) and stay on the bus until alighting at stop 1553 and at stop 10142 (near the
left end of the graph). The route consisted of LOS F segments between stops 5559 and 1553, which are the “pinch points” before the re-routing process was applied. After the SORTAH algorithm was applied, the overcrowded segments were relieved of congestion and were within its capacity (i.e. LOS E). This is an indication to TTC that they can consider developing and deploying a form of policy or strategy on the short to medium-term that can drive network effects towards a user-constrained system optimal solution. As a result, this route could be relieved of congestion. However, TTC should be cautious that as ridership increases in the future, these policies or strategies alone cannot relieve congestion. An increase in capacity or an additional relief route may be required to accommodate the increased demand and eliminate overcrowding on the critical routes.

**Figure 19. Passenger Load Factor Profile and LOS Standards of Route 102 (Southbound) for the Before and After Re-Routing Process from 7:00 am to 8:00 am**
Figure 20 shows a passenger load profile diagram for the same period (7:00 am to 8:00 am), but for Northbound Route 102. The graph shows that not as many people are using the bus route to go north. This is reflective of the real-world, in which people typically go to work near the downtown area of Toronto during the AM peak period. The loads would be reversed during the PM peak period when passengers are commuting from work to home. In other words, there would be more passengers travelling northbound than southbound during the PM peak. In the “before re-routing” scenario, it has been found that the route consisted of at most, LOS D segments. After the SORTAH algorithm was applied, many passengers were re-routed to this route. As shown, the majority of passengers board the train between stops 5557 and 5561, in which the load is at capacity. The resulting segments had a grade of at most LOS E. In the user-
constrained system optimal solution, this route would be functioning at capacity. SORTAH spreads out the demand to better make use of routes that are underutilized.

6 Conclusion

In conclusion, the TTC network still experienced congestion after applying the SORTAH algorithm. The user-constrained system optimal solution can be achieved by minimizing crowding levels, but the re-routing process does not completely eliminate congestion and overcrowding. SORTAH relieves many of the routes of heavy congestion and reduces the number overcrowding segments by spreading out the demand to better utilize routes that are under capacity. Overcrowding cannot be eliminated in a user-constrained system optimal solution because passenger travel behaviour is constrained by the design of the transit network. Passengers realistically consider between four to seven alternatives for travel choices and typically take paths with minimal transfers, minimal travel time and maximal comfort. Given the network design, passengers may be constrained to always travel through a section of a specific route for all path choices, because alternate routes may cause additional transfers, significant travel time, or discomfort from overcrowding for the passenger. Thus, realistic passenger paths are bounded by the network design. In such cases, SORTAH is unable to spread out the demand to other underutilized routes. Additional capacity or a relief route should be considered to accommodate the demand and eliminate overcrowding on the critical routes.
Chapter 7
Conclusion

1 Chapter Overview

This chapter summarizes the key findings of this thesis and provides future directions to advance heuristic application to transit assignment and the “reverse engineering” framework for policy and strategy development.

- Section 2 provides a summary of the key findings and an evaluation of the SORTAH method.
- Section 3 outlines the future work in extending the SORTAH capabilities and SORTAH application.

2 Key Findings

2.1 SORTAH Prototypes

The goal of the thesis is to solve for the user-constrained system optimal solution in which the overall congestion effects of a transit network is minimized while individual travel needs are accounted for. Three variations of the SORTAH prototype were developed to test the hypothesis that total travel time will indirectly be minimized when crowding levels are directly minimized. Each variation of the SORTAH prototype differed by the target travel time criterion implemented. They ranged from the strictest to relaxed travel time criterion:

Case 1 – Shorter Time Check: Target path has smaller travel time than original path.

Case 2 – Measure of Unfairness: Target path has smallest measure of unfairness.
Case 3 – No Time Check: Travel time criterion is relaxed.

The results indicated that the use of Case 1 (Shorter Time Check) does not perform as well as the use of Case 2 (Measure of Unfairness) or Case 3 (No Time Check) in terms of the following elements of analysis:

- Crowding levels – Measured by the number of LOS F segments.
- Total travel time – Measured by the network total travel time.
- Execution time – Measured by the time it takes to run the prototype, relative to the network size.

### 2.1.1 Crowding Levels

When passengers were re-routed to a path with a shorter travel time, the number of overcrowded routes increased while the number of overcrowded segments decreased. This is an indication that Case 1 (Shorter Time Check) spreads out overcrowding amongst the network routes rather than eliminating as many overcrowded routes as possible. When passengers were re-routed to a path with the smallest measure of unfairness, both the number of overcrowded routes and number of overcrowded segments decreased. The overall congestion on the transit network had decreased by eliminating as many overcrowded routes as possible. Similarly, when travel time criterion was relaxed during the re-routing process, the number of overcrowded routes and number of overcrowded segments decreased. Case 2 (Measure of Unfairness) and Case 3 (No Time Check) improved operational efficiency with respect to comfort and crowding levels.

### 2.1.2 Total Travel Time

When passengers were re-routed using either re-routing strategies, the change in average passenger travel time and total travel time had little impact on the passenger travel experience. Case 1 (Shorter Time Check) that used a stricter travel time criteria saved each passenger an average of 6 seconds per trip. Case 2 (Measure of Unfairness) and Case 3 (No Time Check) increased an average of 12 seconds and 30 seconds, respectively, per trip. These changes in
travel time have little impact on passenger’s travel experience because they are too small for passengers to notice a difference. Passengers would start noticing changes in their travel time when the marginal travel time is two minutes or more. Thus, total travel time alone is not a good measure of operational efficiency. Though Case 2 (Measure of Unfairness) and Case 3 (No Time Check) increase the travel time, the use of the measure of unfairness as the travel time criterion is preferred because it minimizes the marginal total travel time. When travel time criterion is relaxed, fluctuations in travel time is not directly controlled like in Case 1 (Shorter Time Check) or Case 2 (Measure of Unfairness).

While these results do not support the hypothesis in which directly minimizing crowding levels would indirectly minimize travel time, it does not disprove the hypothesis. SORTAH currently analyzes the network on a segment basis; therefore, the total travel time is an estimate and does not capture all network effects. Travel time in the model is given on a passenger-by-passenger basis and summed up to yield the total travel time. When re-routing occurs, only the passenger who has been re-routed has their travel time updated. Impact on passengers that use the same segments are not captured due to the limitation of the model that does not calculate travel time on segment-basis. When improvements on SORTAH are available to capture all network effects in the network’s total travel time, the hypothesis can be further examined.

2.1.3 Execution Time

Real-world application of SORTAH requires the heuristic to find a good solution to assist planners in making a quick and well-informed decision on policy implementation. The execution time measure evaluated how fast each variation of the prototype can solve for the user-constrained system optimal solution. Results showed that Case 3 (No Time Check) had the fastest execution time of less than one hour. This is because the travel time criterion was relaxed and there were fewer criteria to check during the re-routing process. Case 1 (Shorter Time Check) or Case 2 (Measure of Unfairness) had comparable execution times which are approximately two times the execution of Case 3 (No Time Check). All three cases had desirable execution time because they did not exceed more than two hours.
2.1.4 The Final SORTAH Algorithm

Figure 21 illustrates the final SORTAH re-routing procedure. When all three variations of prototype were compared by the three elements of analysis, it has been found that the SORTAH procedure that re-routed passengers following the algorithm of Case 2 (Measure of Unfairness) had provided better operational efficiency with respect to comfort, crowding levels and travel time. Its execution time was no more than two hours and quickly solved for the user-constrained system optimal solution that minimized crowding levels and marginal travel time. The final SORTAH algorithm was used to further analyze the TTC network as a real-world application.
2.2 Case Study: TTC Network

The case study results indicate that SORTAH does not completely eliminate overcrowding in the TTC network. The SORTAH algorithm relieves as many heavily congested routes as possible and reduces the number of overcrowded segments by spreading out the demand to better utilize routes that are under capacity. Overcrowding cannot be eliminated in a user-constrained system optimal solution because passenger travel behaviour is constrained by the design of the transit network. Passengers realistically consider between four to seven alternatives for travel choices and typically take paths with minimal transfers, minimal travel time and maximal comfort. Given the network design, passengers may be constrained to always travel through a section of a specific route for all path choices, because alternate routes may cause additional transfers, significant travel time, or discomfort from overcrowding for the passenger. Thus, realistic passenger paths are bounded by the network design. In such case, SORTAH is unable to spread out the demand to other underutilized routes. Additional capacity or a relief routes should be considered to accommodate the demand and eliminate overcrowding on the critical routes.

When the network was analyzed on a route-by-route basis, it was found that 34 out of 148 routes were overcrowded in the user equilibrium solution. After applying the SORTAH, 10 out of the 34 overcrowded routes were completely relieved of congestion, but 24 out of the 34 overcrowded routes still contained some overcrowding sections. This is an indication that there are no existing alternate routes in which passengers can be re-routed to. Even with the deployment of ITS technologies that can influence passenger path choice towards a user-constrained system optimal solution to improve operational efficiency, passengers cannot be re-routed to an alternate path. Capacity would need to be added on the heavily congested routes, or a relieve route would need to added to the existing network to accommodate the demand and eliminate overcrowding on the critical routes.
3 Future Work

3.1 SORTAH Extensions

The following future research items can be conducted to resolve SORTAH’s limitations and to improve the SORTAH algorithm:

- **Improve SORTAH to capture all network effects in the network’s total travel time.** Currently, the total travel time is an estimate and does not capture all network effects, because travel time in the model is given on passenger-by-passenger basis and summed up to yield the total travel time, while the network is analyzed on a segment-basis. Perceived travel time factor related to the load factor LOS cannot be applied because route segment travel time cannot be extrapolated from the passenger information. When re-routing occurs, only the passenger who has been re-routed has their travel time updated. Impact on passengers that use the same segments are not captured due to the limitation of the model that does not calculate travel time on a segment-basis.

- **Extend SORTAH to incorporate temporal shifting within the discretized analysis period (i.e. one-hour time period base, such as the period between 7:00 am and 8:00 am within the AM peak period).** For the same path, passengers are restricted to the specific departure times defined for the path. Although multiple transit vehicles can serve the path within the hour, passengers are modelled to depart and board the specific transit vehicle defined. For example, if a path that contains a route with 5-minute headways is modelled to serve the passenger at 7:15 am, the passenger would not be modelled and shifted to board the 7:10am or 7:20 am vehicle. Passengers are assumed to be on time so the transfers of the paths are retained.

- **Consider passenger’s departure time choice.** Currently, departure time choice is ignored in SORTAH. It is assumed that passengers typically go from their origin to their destination within the same hour. They cannot be rerouted to another route of a different hour. For example, if the passenger is modelled to take a path between 6:00 am and 7:00 am, all 7 paths have departure times within the 6 AM hour-base. However, this may not
be true because passengers may experience better travel times and comfort level for departing if he or she departed one hour later, which is outside of the 6 AM hour-based.

- **Account for variability in the arrival of transit vehicles at stops or stations.** Currently, transit vehicles are assumed to always be on-time, which does not capture true network effects. Transit networks typically have schedule adherence and on-time performance concerns that impact passenger travel experience. Transit vehicles that arrive at a stop/station outside of the scheduled time can cause passengers that are waiting for another transit vehicle to board the current vehicle and result in an influx of passengers bunching onto the same transit vehicle. As a result, this may cause a series of overcrowded segments.

### 3.2 SORTAH Applications

In addition to extending the SORTAH algorithm, future work could also be conducted to examine the application of SORTAH:

- **Test for policies that can achieve a constrained system optimum that reduces transit congestion.** The traditional framework for evaluating policies and strategies is to first develop a policy or strategy and then test for its impact via a transit assignment model. The model outputs performance measures that compare the before and after scenarios against the goal of the policy. Depending on the outcome of the after-scenario, transit operators and planners would determine whether the policy or strategy would improve the current performance of the network. With this framework, transit operators and planners cannot determine whether their efforts are well justified when developing the policies and strategies. The motivation is to enhance the current state-of-the-art transit assignment model to determine the upper bound for which network performance can no longer improve and demand cannot be redistributed any further to relieve the congestion. Therefore, transit operators and planners can better concentrate their efforts on developing effective policies and strategies. This supports the proposed “reverse
engineering” framework for evaluating policies and strategies whereby a network solution would drive the development of a policy or strategy.

- **Examine technological effects on passenger route choice.** Emerging information, communication, sensor technologies and innovative transit operations control strategies are becoming critical elements of viable, competitive public transit systems. ITS technologies have significantly expanded the range of information available to the traveller through ATIS and improved the on-time performance of transit vehicles through the use of APTS. Passengers path choice can be influenced by ATIS, which can result in a user-constrained system optimal solution if deployed correctly. The SORTAH solution can be used as the reference and upper bound to examine technological effects on passenger route choice to develop an ATIS that drives towards a user-constrained system optimal solution. This supports the proposed “reverse engineering” framework for developing ITS technologies.

- **Extend SORTAH into a route guidance system for transit mode.** Currently, there are route guidance systems for auto mode that provide drivers with different path alternatives, typically assuming user equilibrium assumptions. There has been recent research that can upgrade route guidance systems that provide system-optimal paths for drivers (Jahn et al., 2005). Existing transit path finders can use SORTAH algorithm as a path of the route searching engine to influence passenger route choice towards a user-constrained system optimal solution. The transit mode route guidance system is a form of ATIS that can influence travel behaviour to drive the network towards a user-constrained system optimal solution.

- **Examine SORTAH applicability on the influence of the physical design of new transit routes.** Realistic passenger paths are bounded by the network design. Passengers typically consider between four to seven alternatives for travel choices and take paths with minimal transfers, minimal travel time and maximal comfort. Given the network design, passengers may be constrained to always travel through a section of a specific route for all path choices, because alternate routes may cause additional transfers,
significant travel time, or discomfort from overcrowding for the passenger. As the TTC results have shown, overcrowding cannot be eliminated in a user-constrained system optimal solution because passenger travel behaviour is constrained by the design of the transit network. SORTAH is unable to spread out the demand to other underutilized routes. Additional capacity or a relief route would need to be considered to accommodate the demand and eliminate overcrowding on the critical routes. SORTAH may be extended to determine and influence the physical design of new transit routes so the user equilibrium solution of the network would be equivalent to the user-constrained system optimal solution.

This study is the first step in achieving the “reverse engineering” framework for policy and strategy development to reduce transit overcrowding. Future work should be conducted to resolve the limitations and improve the SORTAH algorithm, as well as extending its application to develop transit policies and strategies.
References


King County Metro Transit. (2010). *King County Metro Transit Performance Measurement*. Seattle, WA: King County Transit Metro Transit.


Wardrop, J. (1952). Some theoretical aspects of road traffic research. *Institute of Civil Engineers Part II*, (pp. 325-378).

