ONE APP, TWO SYSTEMS

How WeChat uses one censorship policy in China and another internationally

By Lotus Ruan, Jeffrey Knockel, Jason Q. Ng, and Masashi Crete-Nishihata

NOVEMBER 30, 2016
RESEARCH REPORT #84
Copyright

© The Citizen Lab

Licensed under the Creative Commons BY-SA 4.0 (Attribution-ShareAlike licence). Electronic version first published in 2016 by the Citizen Lab. This work can be accessed through https://citizenlab.ca/2016/11/wechat-china-censorship-one-app-two-systems/.

Document Version: 1.0

The Creative Commons Attribution-ShareAlike 4.0 license under which this report is licensed lets you freely copy, distribute, remix, transform, and build on it, as long as you:

- give appropriate credit;
- indicate whether you made changes; and
- use and link to the same CC BY-SA 4.0 licence.

However, any rights in excerpts reproduced in this report remain with their respective authors; and any rights in brand and product names and associated logos remain with their respective owners. Uses of these that are protected by copyright or trademark rights require the rightsholder’s prior written agreement.

Suggested Citation

Acknowledgements

We are grateful to Zubayra Shamseden (Uyghur Human Rights Project) for translation assistance and Professors Ron Deibert and Jedidiah Crandall for supervision.

About the Citizen Lab, Munk School of Global Affairs and Public Policy, University of Toronto

The Citizen Lab is an interdisciplinary laboratory based at the Munk School of Global Affairs and Public Policy, University of Toronto, focusing on research, development, and high-level strategic policy and legal engagement at the intersection of information and communication technologies, human rights, and global security.

We use a “mixed methods” approach to research that combines methods from political science, law, computer science, and area studies. Our research includes investigating digital espionage against civil society, documenting Internet filtering and other technologies and practices that impact freedom of expression online, analyzing privacy, security, and information controls of popular applications, and examining transparency and accountability mechanisms relevant to the relationship between corporations and state agencies regarding personal data and other surveillance activities.
## Contents

**Key Findings** | 6  
---|---  
**Introduction** | 6  
**Legal and Regulatory Environment** | 8  
**How WeChat Censors Keywords** | 10  
  - Previous Examples of WeChat Keyword Censorship | 11  
  - WeChat Public Account Censorship | 12  
**Tracking Censorship on WeChat** | 14  
  - One-To-One Chat Censorship | 14  
  - Group Chat Censorship | 15  
  - Censored Keyword Content | 17  
    - Event | 18  
    - People | 19  
    - Political | 19  
    - Social | 20  
  - Keyword Censorship Updates | 20  
  - URL Filtering | 22  
**One App, Two Systems** | 26  
**Silent Censorship** | 28  
**The Larger the Audience, the Greater the Censorship** | 29  
**Motivations for URL Filtering** | 30  
**Conclusion** | 31  
**Data** | 32
Key Findings

› Keyword filtering on WeChat is only enabled for users with accounts registered to mainland China phone numbers, and persists even if these users later link the account to an International number.

› Keyword censorship is no longer transparent. In the past, users received notification when their message was blocked; now censorship of chat messages happens without any user notice.

› More keywords are blocked on group chat, where messages can reach a larger audience, than one-to-one chat.

› Keyword censorship is dynamic. Some keywords that triggered censorship in our original tests were later found to be permissible in later tests. Some newfound censored keywords appear to have been added in response to current news events.

› WeChat’s internal browser blocks China-based accounts from accessing a range of websites including gambling, Falun Gong, and media that report critically on China. Websites that are blocked for China accounts were fully accessible for International accounts, but there is intermittent blocking of gambling and pornography websites on International accounts.

Introduction

WeChat, (Weixin 微信 in Chinese), is the dominant chat application in China and fourth largest in the world, with 806 million monthly active users.

WeChat encompasses more than just text, voice, and video chat; it includes a rich set of features such as gaming, mobile payments, and ride hailing, which make it more of a lifestyle platform than a mere chat app. It is estimated that Chinese users spend a third of their mobile online time on WeChat and typically return to the app ten times a day or more. WeChat is owned and operated by Tencent, one of China’s largest technology companies.

Operating a chat application in China requires following laws and regulations on content control and monitoring. Accordingly, the popularity of WeChat has also been met with suspicions of surveillance and media reports of censorship. Despite these concerns, there is limited technical research into the operation and scale of content monitoring and filtering. In this report, we provide the first systematic analysis of keyword censorship and URL filtering on WeChat to determine how the app filters content and the type of content that is blocked.
We found that keyword filtering is enabled on WeChat for users with accounts registered to mainland China phone numbers. Filtering remains enabled even if users later link their account with a non-mainland China number, which means that users with accounts registered to mainland China will remain under censorship regardless if they travel or unlink their Chinese phone number from the account. Differentiating content access based on user registration seemingly creates a “one app, two systems” model of censorship.

WeChat performs censorship on the server-side. When you send a message it passes through a remote server that contains rules for implementing censorship. If the message includes a keyword that has been targeted for blocking, the message will not be sent. Documenting censorship on a system with a server-side implementation such as WeChat’s requires devising a sample of keywords to test, running those keywords through the app, and recording the results.

We used a sample of keywords found blocked on other apps used in China and systematically tested that sample in two modes: one-to-one chat and group chat. We found a greater number of keywords blocked on group chat compared to one-to-one chat, which suggests that communications on group chat are specifically targeted, potentially because group chats can reach a larger number of users.

In both chat modes, users are no longer presented with a warning message when they enter blocked keywords, as indicated by previous reports. This change means there is no feedback to users that censorship has occurred making the restrictions on WeChat less transparent. Censored keywords spanned a range of content, including current events, politics, and social issues.

In addition to keyword censorship, WeChat implements a URL filtering system in its built-in browser, which uses different lists of blacklisted and whitelisted websites for China and International accounts. To sample which URLs WeChat censors, we used a script to automatically test the Alexa Top One Million list of websites using both China and International accounts.

We found that 41 of the websites we tested blocked only on accounts registered with mainland Chinese phone numbers. Moreover, every site that is uniquely blocked on China accounts is fully accessible on International accounts, meaning that international users can successfully access the same URLs with WeChat’s internal browser. However, we did find intermittent blocking of other gambling and pornography websites on International accounts.
We proceed by providing an overview of the legal and regulatory system in China, past work on censorship on WeChat, report our new results, and conclude with a discussion on the implications of our findings.

**Legal and Regulatory Environment**

WeChat thrives on the huge user base it has amassed in China, but the Chinese market carries unique challenges. Any Internet company operating in China is subject to laws and regulations that hold companies legally responsible for content on their platforms. Companies are expected to invest in staff and filtering technologies to moderate content and stay in compliance with government regulations. **Failure to comply** can lead to fines or revocation of operating licenses. In 2010, China’s State Council Information Office (SCIO) published a major government-issued document on its Internet policy. It includes a list of prohibited topics that are vaguely defined, including “disrupting social order and stability” and “damaging state honor and interests.” In late-May 2014, China’s State Internet Information Office (SIIO), Ministry of Public Security (MPS), and the Ministry of Industry and Information Technology (MIIT) jointly launched a month-long campaign targeting Chinese instant messaging (IM) services in a bid to clean up “illegal and harmful information” and to fend off “hostile forces at home and abroad.”

In recent years, WeChat has faced increased regulatory pressures. WeChat offers a microblogging feature called “Public Accounts” that allows certain users to publish daily posts. On March 13, 2014, Tencent shut down nearly 40 Public Accounts without giving any prior notice. Popular Public Accounts that discuss current affairs and politics, such as the Consensus Website (共识网), Truth Channel (真话频道), Luo Changping (罗昌平), and Elephant Magazine (大象工会), were shut down overnight. Tencent issued a statement explaining that it “strictly prohibits publishing pornographic, vulgar, violent, bloody, political rumors and any illegal content.” The company said the action was “part of the commitment to providing quality user experience on Weixin in China,” and that it would “continually review and take measures” on suspicious content.

In August 2014, the SIIO announced rules on instant messaging tools, requiring service providers to obtain “Internet news service qualifications,” users to authenticate their identities before registering, public accounts owners to undergo “examination and verification” by the companies, and store this information on file with the “controlling department for Internet information and content.”
This strict regulatory environment has led to suspicions that communications on WeChat may be monitored. There have also been cases of Tibetans being arrested for sharing chat messages, songs, and photos on WeChat with content related to the Dalai Lama and Tibetan culture that Chinese authorities alleged carried “anti-China” sentiments.

Beyond the Chinese market, WeChat has made considerable efforts to grow its user base internationally. Tencent launched advertising campaigns targeting foreign markets, recruiting football star Lionel Messi and Bollywood actors to endorse the app. However, the impact of these efforts has been questionable. Tencent has never disclosed how many active users it has outside of China, but WeChat has yet to make the same impact in other countries as it has in its home market. Some commentators speculate that WeChat has not enjoyed the same success internationally because outside of China the application does not have the same rich set of features, such as mobile payments and taxi hailing, that make it a compelling platform for users within China.

Market growth outside of China has also been hampered by incidents that remind international users of the restrictions WeChat faces at home. In January 2013, media reported that WeChat users outside of China experienced censorship of chat messages that contained the keywords “法轮功” (“Falun Gong”) or “南方周末” (“Southern Weekend”), a Guangzhou-based liberal newspaper in China (see Figure 1). Tencent responded with a statement that claimed a technical error had enabled keyword filtering for international users temporarily and that immediate actions would be taken to rectify the issue.

Figure 1: Screenshots from media reports show international users experiencing keyword censorship on WeChat. Source: Tech in Asia and The Next Web
In 2015, WeChat introduced a temporary feature to commemorate Martin Luther King Day in the United States. If users typed “civil rights” into the chat window, animated American flag emojis would rain down on the screen (see Figure 2). This feature was only intended for users based in the U.S., but was accidentally enabled for China-based users. Tencent was criticized in China for the mistake, quickly disabled the feature for China based users, and issued a statement: “WeChat’s path to internationalization isn’t easy… We will try even harder!”.

These incidents demonstrate the balancing act Chinese tech companies must perform, as they attempt to grow outside of China while staying within the lines of domestic regulations.

Figure 2: Screenshot of animated American flags raining down on the screen when WeChat users typed “civil rights” into the chat window.

**How WeChat Censors Keywords**

Keyword censorship can be implemented in two ways: on the client-side (i.e., on the application itself) or on the server side (i.e., on a remote server). In a client-side implementation, all of the rules to perform censorship are inside of the application running on your device. Often the application has a built-in list of keywords that it uses to perform checks to determine if any of these keywords are present in your chat messages before your messages are sent. If your message contains a keyword from the list then the message is not sent. In a server-side implementation the rules to perform censorship are on a remote server. When a message is sent, it passes through the server that checks if banned keywords are present and, if detected, blocks the message.
Client-side implementations can be analyzed by reverse engineering the application and extracting the keyword lists used to trigger censorship. A censorship keyword list provides a comprehensive look into exactly what content an application was censoring over a specific period of time. Previous research has uncovered client-side censorship in **TOM-Skype** (the version of Skype available for the Chinese market until 2013), **Sina UC** (a chat app that was provided by Sina Corporation), and **live-streaming platforms** used in China. Client-side censorship was also found in **LINE**, a mobile chat client developed by a Japanese company and marketed to countries around the world, including China. The keyword filtering features in LINE were only enabled for users with accounts registered to mainland China phone numbers in an effort to comply with Chinese regulations.

Server-side implementations pose greater challenges for researchers. Analyzing server-side implementations generally rely on sample testing in which researchers develop a set of content suspected to be blocked by a platform, send the sample to the platform, and record the results. In the case of a chat app, this process means developing a set of keywords suspected to be blocked, sending these keywords in a chat, and documenting if the keyword is received or not and if any warning message is presented. In comparison to extracting keyword lists from client-side implementations, sample testing cannot gain a comprehensive view of what a platform is censoring, as the results are only as accurate as the overlap between the sample and the actual content filtered.

WeChat censors keywords on the server-side. Therefore, sample testing has to be used to determine what specific keywords are blocked. In the next section, we describe previous sample testing results on WeChat.

### Previous Examples of WeChat Keyword Censorship

Following the 2013 media reports of international users experiencing keyword filtering, we ran a series of tests to attempt to document the presence of the filtering and determine the conditions that trigger it.

In May 2013, using a WeChat account registered to a U.S. phone numbers while on a Chinese network, we found keyword censorship of “法轮功” (Falun Gong) but not for “南方周末” (Southern Weekend). Running the same test from a U.S.-based network with the same accounts resulted in no censorship for “法轮功”. These results suggest that at that time, censorship was triggered depending on what network the user was on.
In December 2013, we ran a similar test using an account registered to a mainland China phone number while on a Canadian network. Again, our test found that the keyword “法轮功” (Falun Gong) was being filtered but “南方周末” (Southern Weekend) was not. Figure 3 shows screenshots from our two rounds of testing in 2013.

![Figure 3: Evidence of censorship in WeChat’s one-to-one chat feature. Left: May 2013, WeChat client using an account registered to a U.S. phone number running from a Chinese network. Right: Dec 2013, WeChat client using an account registered to mainland China phone number running from a Canadian network.](image)

In tests undertaken later, in January and February 2014, we were unable to reproduce the blocking of “法轮功” (Falun Gong) nor did we uncover any censorship when we tested sets of keywords extracted in previous work on chat app censorship. We were also unable to trigger blocking when attempting to reproduce the conditions of our May 2013 test by using a VPN based in China and by spoofing GPS locations in China.

**WeChat Public Account Censorship**

In 2012, WeChat introduced a Public Accounts platform (微信公众平台), which allows individuals and companies to publish short blog posts to which other users can subscribe.

In a previous Citizen Lab report, Jason Q. Ng provided the first attempt at systematically identifying what is censored on the Public Account platform by downloading over 36,000 unique public account posts between June 2014 and
March 2015, monitoring them over time and tracking whether they were deleted, distinguishing between posts that the system reports were deleted by the user or posts that the system acknowledges were censored by WeChat. The results suggest that there is a list of blacklisted keywords—possibly including words related to Falun Gong and June 4 (the anniversary date of the Tiananmen Square protests), among others—that would set off automatic review filters, which prevent a post from ever being published in the first place. Furthermore, posts containing keywords related to issues such as corruption and Chinese officials were much more likely to be censored by WeChat.

Also of note was the error message WeChat provided indicating that a post had been censored (see Figure 4). The reason for the deletion is attributed to a WeChat users’ peers, suggesting that WeChat is playing the role of a hands-off moderator, letting its users independently decide whether a piece of content is appropriate and only acting as a judge after a post has been flagged too many times. However, analysis by Jason Q. Ng shows that certain sensitive keywords were vastly underrepresented in both censored and uncensored posts. This finding combined with anecdotal reports by Public Account users indicated the presence of built-in automatic review filters that preemptively blocked posts from ever being published. Furthermore, a thorough reading of hundreds of censored posts also raises doubts about how many were organically reported by users as opposed to actively removed by WeChat. This questionable framing of WeChat as a neutral third-party in censorship decisions is similarly seen in the error messages provided when a website is blocked from access in WeChat’s internal browser. This issue of attribution and transparency of censorship (or lack of it) are issues explored in the Discussion section of this report.

Figure 4: The warning message presented when a user accesses a censored WeChat Public Post reads, “The content has been reported by multiple people. Relevant content cannot be viewed.”
Tracking Censorship on WeChat

In June 2016, we found that the keywords “法轮功” (Falun Gong in simplified Chinese) and “法輪功” (Falun Gong in traditional Chinese) were filtered in one-to-one chat on WeChat on an account registered to a Chinese phone number used on a Canadian network. In our previous tests, users were sent the warning message: “Your message could not be sent due to local laws, regulations, and policies.” In the June 2016 tests, no warning is sent to the user and the message is not delivered to the receiver. Neither party is aware that any censorship has taken place. Unless the sender and the receiver compare their chat logs, there is no indication that the message was not delivered or received. We found that “法轮功” is similarly censored in group chat with no notifications provided to the sender nor to any of the other members in the group.

Following these initial findings, we ran systematic tests to determine the content, scope, and triggers of keyword censorship in one-to-one and group chats. We registered four accounts for testing purposes: two registered to mainland China phone numbers, one to a Canadian phone number, and one to a U.S. phone number. We then conducted sample testing using keywords we had previously extracted from other applications used in China that implement client-side keyword filtering.

One-To-One Chat Censorship

In June 2016, we tested 26,821 keywords from our keyword sample using two accounts registered to mainland China phone numbers on a Canadian network. Out of our sample, only “法轮功” (“Falun Gong” in simplified Chinese characters) and “法輪功” (“Falun Gong” in traditional Chinese characters) were filtered (see Figure 5).

Figure 5: Evidence of censorship in WeChat’s one-to-one chat feature. The users cannot send or receive the message with the keyword “法轮功”. No indication is provided to either user that the message has been blocked.
In August 2016, we ran the same tests using two accounts, one registered to a mainland China number and the other to a U.S. number. We conducted all our tests on Canadian networks. We could not reproduce the filtering results of “法轮功” or “法輪功” and did not find any other keywords blocked from our sample. Figure 6 shows a China account successfully sending “法輪功” to a Canadian account.

![Figure 6: In August 2016, our tests showed that “法輪功” was not filtered in one-to-one chats.](image)

**Group Chat Censorship**

WeChat offers a group chat feature that allows up to 500 users to share a chat room.

In July 2016, we tested the same 26,821 keywords on group chat using four accounts (two registered to China numbers, one to U.S., and one to Canada). For these tests, we used the China accounts as the designated message sender.

Initially, we tested a large number of keywords from our list at a time by copying and pasting into the chat window a few hundred keywords at once. If the list was censored, we would then use binary search to test half of the list at a time to narrow in on the blocked keyword. A similar methodology has been used by researchers to map out keywords used by China’s national level web filtering system, commonly known as the the Great Firewall.

Our testing quickly found numerous instances in which combinations of keywords would trigger blocking but if the same keywords were sent individually they would not be blocked. Figure 7 shows an example of this blocking. When the user with a China account sends a message with the keywords “六四” (six four), “学生” (student), and “民主运动” (democracy movement), the message is blocked, but if these keywords are sent individually the message goes through. The three
keywords will trigger censorship if they are all present in any part of the message as shown in our example: “我的生日在六四。我是一名学生。我在读一本关于民主运动的书” (“My birthday is on June 4. I am a student. I’m reading a book about democracy movements”).

Other keywords must be used as a phrase to trigger blocking. In Figure 8, a user with a China account attempts to send a message with the phrase “六四纪念馆” (June 4 Memorial), and the message is blocked. However, if the keywords are separated and used in a sentence the message is received: “今天是六四，我要去纪念馆” (“Today is June fourth, I will go to the memorial”).

Figure 7: A user with a China account attempts to send a message with the keyword combination “六四” [+ “学生” [+ “民主运动” in a group chat to users with International accounts and is blocked. If the keywords are sent individually they are received.

Figure 8: A user with a China account attempts to send a keyword phrase “六四纪念馆” in a group chat to users with International accounts and is blocked. When the keywords were separated, the message was received.
To streamline the testing, we wrote a python script that takes our sample list of 26,821 keywords and partitions it into segments of keywords, each at most 100 keywords. The script starts at the beginning of the list and generates a segment by sequentially adding keywords from the list until either the segment is 100 keywords long or adding another keyword would cause the segment of keywords when pasted into WeChat to be censored. After the first segment is complete, the script continues using keywords from the list to begin building the second segment using the same rules, and so on until every word from the list is partitioned into a segment. Using this method, if we have already discovered every keyword censored on WeChat, then none of the keyword segments generated will be censored when we try sending them. This method allows us to avoid time unintentionally triggering censored keywords that we had already discovered but that were hidden in combinations of other keywords in ways that are difficult to recognize manually.

**Censored Keyword Content**

Out of our keyword sample we found 174 keywords that trigger censorship. These keywords include 95 keyword phrases and 79 keyword combinations. Table 1 shows the distribution of languages and keyword types.

<table>
<thead>
<tr>
<th>Language</th>
<th>Keyword Combo</th>
<th>Keyword Phrase</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simplified Chinese</td>
<td>60</td>
<td>79</td>
</tr>
<tr>
<td>Traditional Chinese</td>
<td>17</td>
<td>11</td>
</tr>
<tr>
<td>Uyghur</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>English</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 1: The number of blocked keywords in different languages per keyword type.

We used machine and human translation to translate the keywords to English and analyzed the context behind each one. Based on interpreting these translations with contextual information, we coded each keyword into content categories grouped under general themes according to a code book we developed in previous work.

**Figure 9** shows the distribution of the themes and categories. We describe each theme in detail below.
Event

The Event theme includes keywords that reference 6 distinct events. The highest percentage of Event keywords and blocked keywords overall were related to the June 4 1989 Tiananmen Square Massacre, which remains one of the most taboo events in China. Reactive censorship on social media in China often accompanies the anniversary of the event, and the government continues to push revisionist narratives of what happened. Previous research on keyword censorship on chat apps and live streaming apps have also found a high percentage of June 4 related keywords blocked. Keywords in this category included 50 keyword combinations (e.g. 真相 [+ 共产党 [+ 64屠杀, “Truth [+ Communist Party [+ 64 Massacre”]) and 46 keyword phrases (e.g. 六四天安门, “June 4 Tiananmen”; 勿忘六四, “Don’t Forget June 4”).

Other event related categories had single keyword references. Events include corruption cases featuring high profile leaders such as Bo Xilai (我的最后陈述 薄熙来, “My last statement Bo Xilai”), and references to pro-democracy movements in China (茉莉花革命 “Jasmine Revolution”).

Other keywords referenced more obscure events. In November 2015, a popular online travel show, “On The Road” was banned in China after an episode in which the hosts visited Kurdish fighters in northern Iraq and flew a drone that filmed ISIS military positions in neighbouring Syria. The hosts claimed the footage was shared with the French Air Force who used it to target bombing raids. Media reports claim the show may have been banned over concern that it could incite retaliation from ISIS. We found the keyword “Hold Fast Kobane Syria” (坚守科巴尼 叙利亚) blocked on WeChat, which is the title of the episode that led to the show being banned.
People

Keywords in the People theme are all references to officials in the Communist Party of China, including 14 keyword combinations and 7 keyword phrases.

Keywords included derogatory references to leaders such as “Jiang Toad” (江蛤蟆), which is a meme started by Chinese netizens likening the appearance of former Chinese President Jiang Zemin to a toad. Another example uses a combination of keywords (习包子) “Steamed Bun Xi” and (習特勒) “Xi-tler” that reference two nicknames for current Chinese President Xi Jinping. The word steamed bun (包子) is used to refer to Xi following the circulation of a photo showing him ordering lunch at a steamed bun shop that was subsequently criticized as a political show. The second nickname is a comparison of Xi to Adolf Hitler. Recently, a Chinese activist was detained by police when he shared plans to wear a t-shirt with “Xi-tler” and “习包子” on it.

Other keywords reference rumors including the false claim that Jiang Zemin had died (江泽民 死了, “Jiang Zemin Died”) and allegations that Bo Xilai and Zhou Yongkang were planning a coup against President Xi (周永康 薄熙来 政变, “Zhou Yongkang Bo Xilai Coup”).

Political

Categories in the Political theme cover a range of issues including criticism of the CPC, democracy movements, religious groups, and ethnic minority groups. The theme includes 12 keyword combinations and 39 keyword phrases.

Keywords criticizing the CPC include general statements (e.g., 推翻中国共产党, overthrow the Chinese Communist Party”) and references to the Tuidang movement started by the religious group Falun Gong in the early 2000s, which criticises the party and encourages members to withdraw from it (e.g., 九评共产党, “Nine commentaries on the Communist Party”). References to Falun Gong itself are also blocked (e.g., 法轮功, Falun Gong).

The government of China maintains tight control over news media especially those owned and operated by foreign organizations. Blocked keywords include names of news organizations that operate outside of China and critically report on political affairs including Epoch Times (大纪元), Radio Free Asia (自由亚洲电台), and Duowei News (多维新闻).

The government also extends strict regulations to the book publishing industry, pushing dissident and tabloid authors to Hong Kong and Taiwan to publish on
sensitive topics. The sale of banned books was highlighted in 2015 when employees of a Hong Kong bookshop specializing in taboo titles went missing, only to later emerge in custody in mainland China. Their disappearances had a chilling effect on publishers in Hong Kong who pulled sensitive titles from their shelves. Blocked keywords include titles of banned books on alleged power struggles in the CPC and general political gossip (e.g., 十九大争夺战, “19th Party Congress Power Fight”) and combinations of keywords (e.g., 老江 [+] 氣殺習大大/老江 [+] 氣殺習大大, “Old Jiang is More Fierce Than Uncle Xi”).

Another target for blocking are ethnic minority groups. Previous studies of social media censorship in China have found that Tibet-related content is routinely censored. On WeChat, we identify four blocked Tibet-related keywords including references to the Tibetan independence movement (自由西藏, “Free Tibet”) and a Tibetan rights group (藏青会, “Tibetan Youth Congress”). These keywords are noteworthy given the incidents of Tibetans being arrested by Chinese authorities for sharing Tibet-related content on WeChat.

We also found references to Uyghur-related issues. These keywords are in the Uyghur language in both Arabic and Roman script. All of the keywords were related to Islam and generally encourage devotion and sacrifice to the faith (e.g., ئاللاھ يولىدا “for the sake of Allah”, دين ئىسلام يولۋاس “faith is Islam”). Previous research on censorship of live streaming apps in China has also found blocked keywords in the Uyghur language.

Social

The Social theme includes a single keyword referencing online downloads of pornographic images (好莱坞艳照门种子, “Hollywood Sex Photo Gate Torrent”). Our testing sample includes thousands of keywords related to prurient interests, drugs, weapons, and gambling that have been found censored on other applications. It is surprising to find only one keyword related to this category of keywords blocked on WeChat.

Keyword Censorship Updates

Based on new events, we performed multiple informal keyword tests following our two periods of systematic testing and found new keywords blocked in response to these events.

In August 2016, residents of Lianyungang in China’s north eastern Jiangsu province gathered to protest over rumors that their city is a planned site of a nuclear power
plant developed by France and China. On August 17, 2016, we found the following keyword combination blocked on group chat: “15日实行” (15 day carry out) [+] “全市大罢工” (City-wide strike) [+] “灌云” (Guanyun) [+] “连云港” (Lianyungang). When we checked this keyword combination again in November 2016, it was no longer blocked.

From September 4 to 5, 2016, the 11th meeting of the G20 was held in Hangzhou, China. In President Xi Jinping’s opening speech he made a gaffe accidentally saying “reduce taxes and make roads easy [to travel on], facilitate commerce, and loosen clothing” (轻关易道，通商宽衣) when he should have said “reduce taxes and make roads easy [to travel on], facilitate commerce and be lenient to farmers” (轻关易道，通商宽农). This slip of the tongue was clearly embarrassing for Xi. State propaganda departments issued orders to media and technology groups to censor references to the gaffe and keywords related to the incident have been found blocked on live streaming apps.

On September 6 2016, we found that the keyword “通商宽衣” (“Facilitate Commerce and Loosen Clothing”) was blocked on both one-to-one chat and group chat (see Figure 10). As of November 25 2016, this keyword remains blocked on both chat modes.

![Figure 10: The user on the left with a China account attempts to send “通商宽衣” in a one-to-one chat to an International account and is blocked.](image)

From October 24 to 27 2016, the Sixth Plenum of the 18th Communist Party of China Congress was held in Beijing. During this party meeting, Chinese President Xi Jinping received a status lift when the Communist Party gave him the title of
“core” leader. It has been over a decade since the title has been used and was only previously given to three leaders: Mao Zedong, Deng Xiaoping, and Jiang Zemin. On November 25, 2016, we found the keyword “习核心” (Xi Core) blocked on group chat.

These cases of episodic blocking show that keyword censorship on WeChat is dynamic and influenced by emerging news events. We have observed similar patterns with microblogs, chat apps, and live streaming platforms used in China, which also reactively censor content in response to events. The shifts in the keywords blocked on WeChat pose challenges for research since documenting these occurrences requires testing the right keywords at the right time.

**URL Filtering**

In our initial testing, we observed that if certain websites are accessed directly via WeChat’s internal browser, a warning message is returned that explains “As monitored by Tencent Security Centre’s Mobile Manager, this website may contain malicious fraud content. Visiting the website has now been terminated.” This warning message is displayed for any blocked URL (see Figure 11).

![Figure 11: Screenshot from WeChat Web presenting a generic warning message for blocked websites "As monitored by Tencent Security Centre's Mobile Manager, this website may contain malicious fraud content. Visiting the website has now been terminated."

On the mobile and desktop versions, WeChat gives more details about why certain web pages are blocked (see Figure 12). In total, we documented eight unique English warning messages and 13 unique Chinese warning messages depending on the type of content being accessed (see the copies of the warnings in text and html).
To automate the testing of URL censorship, we analyzed the web version of WeChat and found that it implements censorship by pointing each link to an interstitial page. Depending on the link, the interstitial page either (1) silently sends the user via an HTTP redirect to the link, (2) warns the user that some pages may be malicious and has a link for the user to manually click on to get to the page, or (3) blocks the page and displays a reason for the block. Most sites belong to case (2), which appears to be the default case. We call sites in cases (1) and (3) “whitelisted” and “censored,” respectively.

Using this methodology, we tested the entire list of Alexa’s Top One Million Websites under different accounts and on the following three different network vantage points:

1. A Canada account on an American network
2. A China account on an American network
3. A China account on a Chinese network

We found that some sites (including pornography and gambling websites) that we found blocked were intermittently accessible. We hypothesize that this may be due
to load balancing and some servers using older versions of the URL lists. Although we found evidence of load balancing (i.e., multiple IP addresses per hostname and multiple TCP timestamp clocks per IP address), we cannot definitively show that load balancing is responsible for this phenomenon.

To eliminate the variability of this phenomenon, we retested every URL that not all three vantage points agreed upon. After any retest, if all three vantage points now agreed on a URL, we eliminated it from further retesting. On the third retest, no more new URLs were in perfect agreement than were after the second retest with one exception.\footnote{There were still 14 URLs blocked only when testing a Canadian account on an American network, but inspection shows that these were inconsistently blocked and so would have eventually been eliminated after enough retests, and so we disregard them from our analysis.} We used this methodology since the aim of our analysis was to compare the differences between censorship on different accounts and networks, and we preferred false negatives (i.e., failing to report differences between vantage points) to false positives (i.e., falsely reporting a difference between vantage points).

**Table 2** provides a summary of our results.

<table>
<thead>
<tr>
<th>Unique to vantage point(s)</th>
<th>Case</th>
<th>Number of unique URLs</th>
</tr>
</thead>
<tbody>
<tr>
<td>China account on either network</td>
<td>Blocked</td>
<td>41</td>
</tr>
<tr>
<td>Canada account on American network</td>
<td>Whitelisted</td>
<td>90</td>
</tr>
<tr>
<td>Either account on American network</td>
<td>Whitelisted</td>
<td>24</td>
</tr>
<tr>
<td>China account on either network</td>
<td>Whitelisted</td>
<td>254</td>
</tr>
</tbody>
</table>

**Table 2:** The number of unique URLs blocked or whitelisted using different vantage points.

**Figure 13** shows an example of a website blocked (ndtv.com) on a China account, being accessible from an international account.
We automatically categorized the URLs collected from the three vantage points using Bright Cloud, a URL classification service. Manual adjustments were made to the automatically categorized results where necessary.

The results of this categorization show that the majority of the websites (76%) blocked only on China accounts are related to online gambling (see Figure 14). The second largest category (20%) was news and media websites, which include Falun Gong supported media (e.g., ndtv.com), websites reporting on human rights issues in China (e.g., peacehall.com), and the website of the International Consortium of Investigative Journalists (icij.org) which reported on the Panama Papers.

Figure 14: Distribution of the categories of blocked websites. A total of 41 websites on the Alexa’s Top One Million Websites list the are uniquely blocked to Chinese WeChat accounts.
We also found that every site that is uniquely blocked using China accounts is uniquely whitelisted for the Canadian account on the American network (view the full list on our github).

As of October 4, 2016, URL filtering appears to have been lifted for all non-China based accounts for the websites in our test sample.

One App, Two Systems

Technology companies operating in multiple jurisdictions face the challenge of complying with varying content regulations. Google, Facebook, Twitter, and other major companies have published Transparency Reports detailing the requests they receive from governments to take down content. Some applications have added features that prevent users from seeing certain content based on their location. For example, Twitter withholds content from users based on where they are currently located if the company has received and approved a government request to block content in that jurisdiction.

Social media companies operating in China are expected to maintain content filtering to comply with government regulations. Foreign companies entering the Chinese market must also take on this cost of doing business. When LINE Corporation attempted to push its chat app into China, it enabled keyword censorship on the app if a user registered an account with a mainland China phone number. WeChat faces a similar but inverted dilemma. As the company grows internationally, it must maintain content controls for users in China and an unfettered experience for those outside.

Similar to LINE, WeChat enables filtering if a user registers with a mainland China phone number. Unlike LINE, even if the account is later paired with a non-mainland China number it is still subject to the same censorship, which means if the user travels outside of China or moves away from the country they remain effectively within the borders of its content control. Users potentially affected by this restriction are vast: students studying abroad, tourists, business travelers, academics attending international conferences, and anyone who has recently emigrated out of China.

There are at least three potential explanations for this restriction: (1) Unintentional development oversight; (2) Intentional design decision due to range of other features only available to China-based users; (3) Intentional design decision to
ensure mainland China accounts are always under filtering. We further describe and assess each scenario below:

It is possible that the restriction is unintentional and due to a development oversight. WeChat has had problems before with regional features; filtering intended for China-based users was inadvertently turned on for international users, emoji features intended only for U.S.-based users were mistakenly turned on for users in China. Maintaining a large code base for millions of users in multiple jurisdictions with varying content conditions is complicated, and bugs in any software project are inevitable. It is therefore plausible that the account restriction is unintentional.

Keyword censorship is not the only unique feature for China-based WeChat users. Other features such as WeChat Wallet, which allows users to connect credit and bank cards to the app and make mobile payments, are widely popular in mainland China and are not currently available for the majority of international users (see Figure 15).\(^2\) It is possible that for sake of simplicity and efficiency, developers kept all China-only features enabled on the app to ensure users do not lose access to them if they travel or move outside the country. Since features like Wallet include financial data there are clear incentives to prioritize continued access. However, bucketing features by user registration would mean keyword filtering remains on as well.

![Figure 15: Screenshots of user profile pages for WeChat accounts registered in the US (left), China (Middle). Only the China account has access to WeChat wallet features. The interface to WeChat Wallet is shown on the right.](image)

---

\(^2\) WeChat has started to roll out WeChat Wallet in new jurisdictions including Hong Kong and South Africa, but it remains most popularly used in mainland China.
Finally, it is possible that the account restrictions are intentionally implemented to ensure filtering is always on for mainland China registered users even if they travel or move outside of the country. The CPC promotes maintaining ideological connections and guidance with Chinese citizens overseas, particularly for students studying abroad. Tencent may be pressured to ensure filtering persists for Chinese users in any location.

We cannot conclusively determine which of these scenarios is true and it is possible there are other explanations that we have not considered. However, whether the motive is intentional or accidental, the outcome is the same: mainland China WeChat users face censorship regardless of where they are in the world.

**Silent Censorship**

One of our most striking findings is that keyword censorship on WeChat is now performed without any notice to users. In media reports and our own previous tests we observed two different warnings that users would receive if they sent blocked keywords (see Figure 16).

![Figure 16: Screenshots of blocked keyword messages reported in January 2013 by The Next Web (on the left) and from our own tests in December 2013 (on the right).](image)

The first warning message reported by media in January 2013, notifies users which keyword is restricted. The second warning message observed in our May 2013 and December 2013 tests does not specify which keyword triggered the blocking, but explains that the message infringed on “local laws, regulations and policies.” While these messages are vague, they provide a level of transparency to users, confirming the message cannot be sent. Now keyword censorship is done without giving a user any indication that messages are blocked, which significantly decreases the level of transparency in the app.

Other social media and chat apps used in China exhibit varying levels of notification when user content is blocked. On many of these apps, when a user sends a message
with a blocked keyword they are presented with a warning notification that explains the message is prohibited. The user on the other end of the communication may also see receive a warning notification or see the blocked message as a series of asterisks (e.g., ****). Other apps including YY and TOM-Skype silently filtered messages with no notification given to the user. Notably both of these apps also conduct keyword surveillance logging censored messages that users send. China’s national level filtering system (the Great Firewall of China) filters websites without sending users a block page or any other obvious indication that the requested content is censored. Instead users receive a network timeout error, which on the surface could appear to be the result of routine networking issues.

While the exact motivations behind this diminished transparency are unknown, filtering keywords without explicit user notice does provide a semblance of plausible deniability and makes censorship on WeChat resemble how the Great Firewall of China operates, which is by design opaque and unaccountable.

The Larger the Audience, the Greater the Censorship

The greater level of censorship on group chat compared to one-one chat may be due to the semi-public sphere feature of online discussion groups, making the latter subject to a higher degree of scrutiny.

Tencent has been very conscious about turning its public account and group chat features into another Sina Weibo, China’s Twitter-like service, where a single post could spread very quickly and widely on the Internet. Group chat is catching on and is one of the top four activities (61.7%) users engage in on WeChat. In July 2014, Tencent increased WeChat’s group size limit from 100 to 500 users per group. But as the group chat feature grows, restrictions on it tighten. Users who want to join a WeChat group of over 100 persons must first register the WeChat Payment feature and link their WeChat account with their personal bank account (see Figure 17). According to a report, international WeChat users also need to verify their identity by connecting their account with a cellphone number. Since messages on group chats have the potential to reach a wide audience, pressures to control content on the feature may be higher.
Figure 17: When accepting an invitation to join a groupchat with over 100 members in WeChat, a warning message appears, “You need to verify your real identity before accepting the invitation. You can go to ‘Me’ --> ‘Wallet’ to link your bank account and finish verification.”

**Motivations for URL Filtering**

The filtering of URLs in WeChat raises the ongoing question of where the line is appropriately drawn between censorship and legitimate moderation of content. Chinese content providers are required by law to restrict access to certain kinds of online material. And while much publicity is given to the censorship of politically-oriented material critical of the government, some of the content which is blocked also relates to non-political affairs that would similarly be blocked on social media platforms internationally.

A good comparison for how WeChat prevents users from accessing certain websites from within its app is Facebook, which similarly offers such behavior as a protective feature. Users trying to click through to malware or malicious scams are presented with a “link shim,” an interstitial screen which warns a user about the site they are trying to access. Similarly, Google Chrome includes a feature called “Safe Browsing” that protect users against malicious sites and extensions.

As with Facebook, WeChat also offers interstitials reminding users to to be wary of sharing their passwords or private financial information; it could be argued that such screens serve a similar sort of protective service. However, the benefits of such a service are muddled when one considers that other links do not receive any interstitial and are fully blocked. Unlike in Google Chrome where users can choose to view the sites and downloads despite warning messages or simply turn off the Safe Browsing feature, WeChat users are unable to proceed within the app even if they explicitly want to. Removing user choice from accessing content would seem to land the URL filtering feature more toward the censorship side rather than the content moderation one.
And while it is fairly trivial to copy and paste the link into an external web browser, bypassing the WeChat URL filtering altogether, the mere fact that WeChat forces a user to jump through such a hurdle would undoubtedly reduce a large amount of the traffic that would under normal circumstances reach the site unimpeded. Does this feature benefit WeChat users, or does it benefit WeChat itself insofar it can claim to authorities that they are doing their part?

Finally, the source of the URL filtering is ambiguous. While the error messages that appear give reasons for the censorship, it is not often clear how the purported reason matches up with the actual content being censored. Furthermore, as discussed previously, the messages appear to imply that the censorship is bottom up (the content was “reported multiple times,” ostensibly by fellow users) as opposed to imposed from the top down. However, if that is indeed the case, how does one go about reporting a site? What is the redress for incorrectly categorized sites? Is the filtering mostly automated or manual? In this case, transparency is more superficial than real.

**Conclusion**

Internet filtering is practiced by states around the world (including both authoritarian and democratic regimes), potentially leading to what has been described as a “territorialization” of the network. WeChat shows a stark example of how the same borders extend onto popular apps.

As companies push to grow products internationally, pressures from governments to remove content or provide user data are inevitable, as demonstrated by the requests documented in annual corporate transparency reports. For China, attention is typically centered on foreign companies attempting to reach into the market and having to decide how to approach the government’s strict content regulations. Recent news of Facebook reportedly working on features to enable geographically-based filtering features in an effort to re-enter China is but the latest example.

The challenge that WeChat faces is the inverse. Tencent is one of China’s biggest tech companies and WeChat is one of the most popular apps in the country. Expanding beyond China and finding wider success means WeChat must maintain its user base at home (while staying within regulatory boundaries) and also present a compelling experience to attract international users. In response to this situation, WeChat has seemingly created a “One App, Two Systems” model for censorship.
The identity a user creates when they register for WeChat determines which system they will be under. This identity is immutable. Regardless of a user’s location, chosen settings, or attempts to change registration details, they will always be locked into the same system. Even if this restriction is not the functionality intended by WeChat, the idea that such a system is possible is a serious concern for future transnational Chinese citizens--some of whom might in the future even become citizens of other nations.

Also of concern is that WeChat’s notices for censoring content have become much more opaque. Users no longer have any form of notification that messages are filtered on chat and the messages that are provided for URL and public post blocking are ambiguous. Furthermore, the nuance between the contexts in which censorship takes place has grown more sophisticated. WeChat turns on censorship only for users registered from mainland China and imposes greater restrictions on those features within the app which allow for the greatest reach. The recognition of the difference between public spaces (public accounts platform) and semi-public spaces (group chat) suggests a desire to integrate censorship ever more tightly and seamlessly into the app depending on the perceived “threat.”

While these controls, combined with more opaque censorship, do on the one hand provide a more friction-less user experience for the vast majority of users, it also shields them from encountering such censorship. By calibrating censorship in such a way, WeChat may hope to either squeeze out the minority who hope to use WeChat as a source for independent news and commentary (as either a broadcaster or consumer) or limit their audience as much as possible.

As content filtering and other controls become more invisible and seamless to users, documenting how these features work through the type of research demonstrated in this report is increasingly important to ensure users can make more informed decisions about the apps they use and the social media experiences upon which they rely.

Data

List of Keywords found blocked on China accounts
List of URLs blocked on China accounts
URL warning messages (text, html)