UNDERSTANDING AND CIRCUMVENTING CENSORSHIP ON CHINESE SOCIAL MEDIA

A Dissertation
Presented to
The Academic Faculty

by

Chaya Hiruncharoenvate

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
School of Interactive Computing

Georgia Institute of Technology
May 2017

UNDERSTANDING AND CIRCUMVENTING CENSORSHIP ON CHINESE SOCIAL MEDIA

Approved by:

Dr. Eric Gilbert, Advisor
School of Interactive Computing
Georgia Institute of Technology

Dr. Keith Edwards
School of Interactive Computing
Georgia Institute of Technology

Dr. Amy Bruckman
School of Interactive Computing
Georgia Institute of Technology

Dr. Gloria Mark
Department of Informatics
University of California, Irvine

Dr. Munmun De Choudhury
School of Interactive Computing
Georgia Institute of Technology

Date Approved: March 29, 2017

To Teachers,

whose dedication, hard work, and sacrifice

have made the World a better place.

ACKNOWLEDGEMENTS

This thesis would not be possible if I didn't have continuous and numerous support from friends and family back home in Thailand and here in the US. I greatly appreciate all their support and encouragement throughout my education career to complete my PhD. Without all of you, I can't imagine where I would be right now.

First, I would like to thank my parents, Sirima Hiruncharoenvate and Chatree Hirunjaroenvej, for their love, support, encouragement, guidance, and teaching that have shaped me into who I am and have helped me through tough times. Thank you for your understanding and the time you sacrifice to encourage me to finish my degree as we all started this journey together 12 years ago on different sides of the world. Without your love, this thesis would not be at all possible. I would like to dedicate this thesis to you: my parents, my first teachers, my supporters, and my cheerleaders. ♥

Equally as important, I would not be able to get through the 5 years of my PhD program without my advisor, Dr. Eric Gilbert. I appreciate all your advice, from the first one you gave me even before I became your student: recommending me to come to Georgia Tech, to your advice with my thesis, PhD career, and beyond. I can't imagine in what stage of the PhD program I would be right now if I didn't make the decision to join the PhD program here. I'm grateful to be under your mentorship. I will heed all your teaching and advice, and I will always look up to you as my role model.

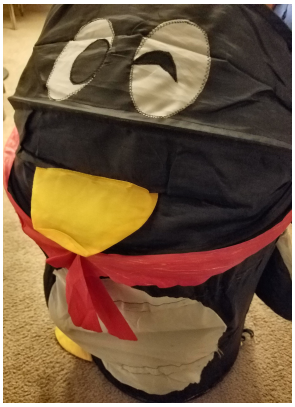
My committee members: Dr. Amy Bruckman, Dr. Munmun De Choudhury, Dr. Keith Edwards, and Dr. Gloria Mark, have provided helpful comments and feedback throughout the course of this thesis, as well as advice for being a good faculty member.

I would like to thank you all my committee members for taking their time out of their busy schedules to help shape my thesis and guide me through the career of academic faculty.

To my labmates: Saeideh Bakhshi, Eshwar Chandrasekharan, Julia Deeb-Swihart, Catherine Grevet, CJ Hutto, Shagun Jhaver, and Tanushree Mitra, I appreciate all your help and support during my time in the comp.social lab. Also, I would like to acknowledge students who have contributed work in this thesis: Mya Havard, Akanksha Mahajan, Jerry Lin, Sam Sherugar, Anurag Shivaprasad, and Kexin Zhang. It was a pleasure working with you. I hope our paths will cross again, and I wish all of you the best in your future endeavors.

Without the Thai community in Atlanta, the past 5 years would have been tougher than what it has been. Thank you P'Lim Pisit Jarumaneeroj for helping me settle in to Atlanta when I first moved here. Thank you P'Nat Prakongpan for being the big brother and taking care of all Thai students here in Atlanta. Thank you the Thai Student Organization and Thai student community at Georgia Tech for your moral support. Keep going and don't lose the fire that you came in with. It does feel really good when you're done. Trust me!

Thanks to all my friends from Thailand and my TS friends for helping me through the past 12 years that I have been in the US. My Debsirin friends (my high school in Thailand), I have kept the gift that you gave me at the airport the day I first came to the US, and I'm still using it until this day (see right ;p). Every time I look at it, it gives me encouragement as if you all were here with me. Thanks to Nina Thanathat Chaiyanon for our frequent exchange of nonsensical and stupid messages that always helps relieve my stress. And finally, thanks to Pae Dolsarit Somseang for making the past 8 years so much better.



Thank you to all my relatives back in Thailand for your support. Since I was young, my relatives have always bet on me to become the first doctor in the family who they can rely on for medical help, but it turns out I'm going to be a different type of Doctor. While this PhD doesn't allow me to treat any of you medically, I appreciate all your encouragement to keep me going, and I hope I have made all of you proud. Don't worry, I have made acquaintances with a bunch of doctors who will be able to take good care of you :)

Thank you to my sponsors: Ministry of Science and Technology and the Royal Thai Government for their financial support, and thank you to the Office of Educational Affairs, Royal Thai Embassy in Washington, DC for their help throughout my education career here in the US. Thank you to the Faculty of Informatics at Mahasarakham University, Thailand for 12 years of patiently waiting for me to complete my Bachelor's, Master's, and PhD degrees. I promise that I will do my best as a faculty member to help educate Thai students to ensure that they receive the best education experience I can give them.

Most importantly, I would like to dedicate this thesis to all the Teachers in my life. In Thailand, *teacher* is not necessarily an occupation but is often a recognition for one's dedication and sacrifice to help others in becoming valuable assets to the society. Thus, the word "teacher" is highly esteemed and well regarded. In my life, I have many Teachers who have taught me valuable lessons to become a good person for the society. I believe that my accomplishment with this thesis will make all of them proud. I heed all your teachings, and hopefully, I will get to repay your kindness in the near future.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SUMMARY

Chinese Internet users not only face the most technologically advanced filtering system in the world, the Great Firewall of China, but also are under the watchful eyes of the repressive government that controls every layer of their communications [19, 126]. Although social networking sites such as Facebook and Twitter are blocked in China, Chinese Internet users have the local replicas such as WeChat and Sina Weibo to communicate with others [116]. However, these sites employ both advanced keyword detection algorithms and human censors to filter any kind of "inappropriate" content [61, 103]. While previous research has explored the technology behind the censorship mechanisms [8, 61], little work has focused on the effects of censorship on online and offline behaviors. In this thesis, I bridge this gap by conducting a mixed-method study to gain a deeper understanding of these effects.

The results of the mixed-method study show that censorship has strong *off-platform* effects, which are not detectable from usage logs. Users deliberately self-censor their speech out of caution, because they do not have a clear understanding of what content is being censored and what risks are associated with censorship on Chinese social media. Although *on-platform* effects of censorship are present on social media usage logs, they wear out over time. Informed by these results, I attempt to provide social media users a better understanding of how the censorship mechanism works and an effective censorship circumvention technique, both of which will lead to greater freedom of expression among social media users.

Digital activists have long employed *homophones* of censored keywords to avoid detection by keyword matching algorithms on Chinese social media [18, 47, 122, 125]. One part of this thesis demonstrates that it is possible to scale this technique

up in ways that are costly and difficult to defend against because human censors must manually read through all social media posts. Specifically, I developed a non-deterministic algorithm for generating homophones that creates large numbers of false positives for censors. In experiments, the algorithm allows homophone-transformed posts to remain on Sina Weibo three times longer than their previously censored counterparts without creating any confusion to native Chinese speakers.

Extrapolating from this work, I employed this algorithm in the development of *CENSE*, a real-time system that Chinese social media users can use to easily detect and replace censored keywords with homophones. The system consists of two primary components: a back-end server and a front-end client. The back-end server handles all logical operations in support of censorship circumvention—extracting censored keywords from Chinese social media and transforming them into corresponding homophones. The front-end client automatically detects censored keywords on users' social media posts and suggest corresponding homophones as replacements. The results of a formative interview study indicate a welcoming response from Chinese social media users to the concept of a censorship circumvention tool.

Overall, the contributions of my research bridge the areas of Internet censorship and censorship circumvention technologies. The mixed-method study provides a better understanding of how censorship affects social media users. Additionally, the homophone transformation algorithm and CENSE, a real-time censorship circumvention tool, aid users in experiencing increased freedom of expression on Chinese social media.

CHAPTER I

INTRODUCTION

The Internet has just recently surpassed newspapers and radios to become one of the most popular sources of news and information [106]. However, not everyone in the world has access to the same information, even though almost half of the world's population has access to the Internet [52]. Because the Internet has facilitated access to information, repressive states across the world have imposed limitations on Internet access on their citizens based on what the regimes regard as appropriate. This thesis takes a closer look at China, the world's most populous country.

With more than 721 million Internet users in China [52], the Chinese government has developed one of the world's most technologically advanced Internet filtering system: the Great Firewall of China. Every packet of Internet traffic in and out of the country is filtered to block specific sites and keywords [19]. While people in other countries spend their daily leisure time on Facebook, Twitter, Instagram, and other popular Western social media, the Great Firewall restricts access to these Western media, and Chinese Internet users frequent local replicas of these social networking sites such as Sina Weibo, Renren, and WeChat as a part of their daily routines [116]. Although Chinese social media appears to contain much content on a variety of topics, the government requires site operators to heavily censor inappropriate content by using both advanced algorithms and human censors [61, 103].

Although previous research has explored the mechanisms underlying this censorship [8, 61] and the ways Chinese social media users employ to circumvent them [18, 47, 122, 125], little is known about the effects censorship has on user's social media

Figure 1: Sina Weibo, the Chinese replica of Twitter, prohibits users to post sensitive content on the site.

behaviors. Therefore, the first part of my thesis bridges this gap in the research literature by examining the user-level effects of censorship on Chinese social media. I conducted a mixed-method study incorporating interviews with Chinese social media users and analysis of Chinese social media user profiles to understand how users modify their online behaviors and contributions to Chinese social media in the presence of censorship.

To combat Internet censorship, Chinese Internet users have employed several techniques when posting to social media, including use of nicknames or morphs of sensitive words and names of political figures, etc. In particular, my focus is on the techniques Chinese Internet users employ to gain increased freedom of publication on Chinese social media. As a part of this thesis, I developed an algorithm that computationally generates homophones of keywords censored on Chinese social media. In experiments conducted to test this algorithm, replacing censored keywords with the homophones generated by the algorithm extended the life of a social media post by three times in comparison to the post containing the original, censored keywords. I then further extended this algorithm into a real-time system that Chinese social media users can use to circumvent censorship and gain a better understanding of the current censorship

situation on Chinese social media.

In this thesis, I organize my work into two categories: understanding the effects of censorship on Chinese social media and circumventing censorship on Chinese social media. Next, I give an introduction of my work in each of the two categories.

1.1 Understanding Censorship on Chinese Social Media

Although the research literature extensively documents the censorship practices on Chinese Internet and Chinese social media [19, 23, 60, 61, 112], we do not know how censorship affects the users. To extend our understanding of these effects, I conducted a mixed-methods study to examine the user-level effects of censorship on Chinese social media. Using the language of distributed cognition [48], I categorized the effects into *off-platform* effects and *on-platform* effects. Off-platform effects, such as self-censorship practices, are not present in user's social media usage log. On the other hand, on-platform effects can be directly observed from the social media profiles such as reduction in subsequent speech and account abandonment after the enactment of censorship.

I conducted both quantitative data analysis and an interview study to explore both on- and off-platform effects of censorship on Chinese social media. In the quantitative data analysis, I analyzed the profiles of more than 1.6 million Sina Weibo users, one of the largest Chinese social media sites. 8,140 of these users were previously censored on the platform. Since the log of social media profiles does not unveil the off-platform effects of censorship, I interviewed 11 Chinese social media users to understand their habits and behaviors in their social media routines.

The results of the matched sampling analysis show weak on-platform effects of censorship. In the 30-day period around censorship, censored users reduced posting activity by 3.91% more than the control group. In comparison to the 5-day period around censorship where censored users' posting activity drops 8.32% more than

the control group, censorship causes short-term suppression of speech. Moreover, in the 30-day period after censorship, the abandonment rate of accounts with censored posts is only slightly more than accounts in the control group: 3.55% censored abandonment rate vs 1.33% control abandonment rate. This implies that although the on-platform effects can be detected, they diminish over time. In contrast, I detected strong off-platform effects. Interview participants reported that they cautiously self-censor around political and sensitive topics due to unclear models of censorship and uncertain associated risks.

The results of this study imply that while both off-platform and on-platform effects certainly exist, on-platform effects are relatively small, and off-platform effects are confined to "sensitive" content. It seems that if Chinese social media users have a better understanding of the current situation of censorship on social media, they will be better informed about what content is appropriate to post on social media, rather than abundantly self-censoring out of caution. I see an opportunity to make the censorship mechanism more transparent to users and, eventually, enable users to circumvent censorship to achieve greater freedom of expression on social media. Next, I introduce the next part of this thesis influenced by the results of this mixed-method study.

1.2 Circumventing Censorship on Chinese Social Media

As documented by [19, 23, 60, 61, 112], the practice of censorship circumvention is common among Chinese social media users. Unlike the English language, certain properties of the Chinese language make constructing words that sound similar or identical to other words, yet have completely different meanings, much easier. Chinese social media users have used this characteristics of the Chinese language in strategies to circumvent censorship on social media [18, 125]. For example, a river crab meme that spread across Sina Weibo did not actually refer to river crabs. Rather, it stood

for a protest against Internet censorship, as the word *harmonize* (和谐, pronounced *hé xié*,), slang for *censorship*, is a homophone of the word for *river crab* (河蟹, pronounced *hé xiè*) [125]. To date, no research has been conducted as to the effectiveness of these strategies in circumventing censorship.

As a part of this thesis, I developed an algorithm that computationally generates homophones of Chinese words and phrases. This algorithm permits the homophone transformation technique to be transformed into a computational algorithm whose implementation can help users circumvent censorship on Chinese social media. To judge the effectiveness of my homophone generation algorithm, I conducted an experiment that compared the life of Chinese social media posts that were transformed using my algorithm with the original posts. In this experiment, the homophone transformation technique extended the life of Chinese social media posts by three times in comparison to the original, unaltered posts. Moreover, in another experiment with Chinese native speakers, the homophone transformation did not confuse them. Furthermore, I analyzed that the homophone transformation is costly for the censorship adversaries to defend against, as automated detection of homophone transformations can cause the censorship algorithm to over-censor regular social media posts.

Extrapolating from this work, I envisioned development of the algorithm into a real-time system that will help Chinese social media users during their daily social media routines. Consequently, I conducted a formative interview study with Chinese social media users to assess whether a censorship circumvention system would be welcomed by such users. The results of the formative study were extremely encouraging, and so I implemented my homophone transformation algorithm in a real-time system, *CENSE* to allow Chinese social media users to apply the algorithm to social media posts in real time.

The system consists of two components: a back-end server and a front-end server.

The back-end system provides components that, working together, continuously monitor for censored keywords on Chinese social media and then computationally generate homophones of these keywords. At the same time, the front-end client unobtrusively monitors the user's post for censored keywords and suggests homophone replacement suggestions when it detects them in the user's post. Overall, the performance of the CENSE system is remarkable, with no lags or latency in user interactions with the system.

1.3 Research Contributions

This thesis contributes to the Social Computing research by providing a better understanding of how censorship affects users on social media, especially in the context of Chinese social media. Consequently, user-level effects can translate to the effects that censorship has on the social media platform. Additionally, this thesis also contributes to the Social Computing Systems research area through the development of an algorithm and a system that enable greater freedom of expression on Chinese social media. More specifically, the contributions of this thesis are as follow:

1. **An understanding of user-level effects of censorship on Chinese social media.** Through a mixed-method study, I found that censorship on Chinese social media cultivates implicit, off-platform effects on users. To put it differently, the effects of censorship do not show through the usage log when analyzing social media user profiles, but users are cautious of censorship when posting on social media. However, the detectable, off-platform effects still exist as my analysis shows a drop in participation right after users are censored, but these effects are ephemeral. While previous research has revealed the mechanisms behind censorship on Chinese social media, the understanding of user-level effects of censorship is still unclear. My work is this thesis presents a unique perspective and complement the research literature with the results of my study.

2. **An algorithm that computationally generates homophones of Chinese words and circumvent censorship on Chinese social media.** I developed a non-deterministic algorithm that computationally generates homophones of Chinese words. Combining this algorithm with a process to detect censored keywords on Chinese social media using TF-IDF, replacing censored keywords in social media posts with their homophones can extend the life of these posts by three times in comparison to the original posts. Furthermore, Chinese native speakers can still understand the content of the transformed posts. Additionally, this technique incurs high cost to adversaries to defend against because human censors are required to inspect these transformed posts manually. Through experiments, my algorithm is proven to be effective in circumventing censorship on Chinese social media.

3. **A real-time system to circumvent censorship of Chinese social media.** Informed by the other parts of this thesis, I developed a real-time system to circumvent censorship on Chinese social media. Besides the homophone transformation algorithm and the censored keyword detection mentioned earlier, I created a social media stream watcher to continuously monitor posts from Sina Weibo. Together, these modules form a back-end server of the system that detects up-to-date censored keywords on Chinese social media and suggests homophone replacements for these words. The other component of the system is a Google Chrome extension as a front-end client. A formative study shows a welcoming response to the idea of an automated censorship circumvention tool from Chinese social media users. To the best of my knowledge, this system is one of the first systems to encourage users to use a homophone substitution technique to circumvent censorship on Chinese social media.

Altogether, this thesis bridges the gap between Internet censorship and censorship

circumvention technologies by exploring how technology changes the perspectives and behaviors of social media users under censorship. The knowledge contributed will help inform future designs of social systems under censorship to minimize the effect of regime-imposed censorship on users and social media platforms.

1.4 Thesis Overview

This thesis is organized into three main parts, each focusing on different study and contributions.

- Chapter 2 presents the related work in the area of Social Computing, Computer Science, and Political Science that informed the work in this thesis. This related work covers the topic of censorship theories, Internet censorship, social media use under repressive governments, and censorship circumvention.

- Chapter 3 presents a mixed-method study including interviews with Chinese social media users and analysis of the profiles of 1.6 million Sina Weibo users to identify the user-level effects of censorship on Chinese social media. This chapter covers the methods used in the study, reports study results, and discusses the results along with design implications based on the study's results.

- Chapter 4 presents the Chinese homophone generation algorithm that I developed and the pipeline of processes that utilizes this homophone generation algorithm to help circumvent censorship on Chinese social media. In this chapter, I also present two experiments I conducted to prove the effectiveness of the censorship circumvention technique.

- Chapter 5 presents an extension of the work in Chapter 4, a real-time system to circumvent censorship on Chinese social media, *CENSE*. Before detailing the design and the development of the two system components, the back-end server and the front-end client, I present the results of a formative interview study

with Chinese social media users regarding their opinions towards censorship circumvention tools. Finally, I present screenshots of a use case scenario of the system designed to circumvent censorship on Sina Weibo.

Finally, I conclude this thesis with a discussion of future directions for the research and suggestions to other researchers and designers in the area of Internet censorship research.

CHAPTER II

RELATED WORK

In this chapter, I present a survey of literature in four areas that have informed and shaped my research: censorship theories, Internet censorship practices, social media use in repressive regimes, and current censorship circumvention techniques.

2.1 Censorship Theories

Internet censorship stems from the control that states have over press and journalists who produce "traditional" media: newspaper, magazines, TV news. Once the Internet gains popularity, states still want to maintain the control of information their citizens receive. However, information on the Internet is harder to control than those on other types of media because national borders are more permeable online; Internet users can easily grab information published in other countries [66]. There are several reasons why states are motivated to control information available to their citizens and, consequently, impose Internet censorship [27, 80, 113, 118]:

- political repression of dissidents, human rights activists, or comments insulting to the states (e.g. China, Iran, Myanmar)
- religious controls (e.g. Arab states)
- protection of intellectual properties (e.g. Denmark, France, Malaysia, Norway)
- cultural restrictions to oppress ethic and sexual minorities (e.g. Indonesia)

In practice, censorship involves control over Internet access, functionality, and contents [30]. Internet censorship consists of three mechanisms: social, political, and technical. Social mechanism put pressure on Internet users not to visit forbidden

sites. However, globalization of information and the ease of connectivity exert a counteracting social pressure in favor for free information. When social mechanism fails, political mechanism applies. Internet users, especially political dissidents who are some of the early users of circumvention technology, are often aware of the threats from political force who are ready to enforce the laws. Technical mechanism acts as both the first and last line of defense against access to undesirable information. In the first place, lawmakers, through technical mechanism, can enable social mechanism by designating which pieces of information citizens should avoid. As the last place, technical mechanism can be triggered by actually blocking when social and political mechanisms fail [126].

As a result, censorship is seen in two forms: direct censorship and self-censorship [85]. Direct censorship involves political and technical mechanisms that the government imposes to explicitly control the information available to their citizens. On the other hand, self-censorship is influenced by social mechanism of censorship to discourage publication of information by private parties. There are several techniques that governments use to restrict and control Internet access, for example [113]:

- harassment of bloggers/whistle-blowers (social)
- tapping and surveillance (social)
- requiring discriminatory ISP licenses (political)
- discriminatory or prohibitive pricing policies (political)
- content filtering based on keywords (technical)
- website blocking of specific IP addresses (technical)
- hardware and software manipulation (technical)
- denial-of-service (DOS) attacks (technical)

Because the original design of the Internet as a distributed system lets the Internet tolerates damages from a single point of failure [66], precise filtering is almost impossible [113]. Thus, states could suffer the risk of overblocking (blocking sites that are not supposed to be censored) [27, 87] and underblocking (allowing sites that are prohibited) [74].

More than 60 countries around the world control their local Internet—some with more restrictions than others [78]. Researchers have been attempting to document censorship practices in specific countries such as Pakistan [75], Iran [5], and China [19, 23, 60, 61]. However, with plethora of practices and techniques each country uses, the task of documenting censorship practices seem to be neverending. While my research does not contribute new knowledge to this area, previous works in censorship theories provide context and shape the methods I use to answer my research questions. In the next few sections, I will show some examples of how different countries around the world control the information on their Internet and the consequences of these policies.

2.2 Internet Censorship

In this section, I review previous works which explore censorship practices around the world. The section is divided into two subsections; the latter focuses solely on works on Chinese Internet censorship since they provide context behind the work of this thesis.

2.2.1 Internet censorship practices around the world

Nearly every country in the world controls the Internet access to their citizens one way or the other [113]. Countries where citizens have more freedom and liberty such as the United States and Australia use more reactive methods such as passing laws to govern the Internet as a way to control information [100]. Other states filter Internet data as it travels onto their local networks. Some countries are more transparent with censorship than others. For example, Saudi Arabian Internet users are notified

when they visit a block page. Internet users can also send requests to block or unblock specific pages [13]. Some states even put the pressure on overseas firms to block access to their own citizens. Twitter routinely releases removal requests from several countries, revealing that the company has received the most removal requests from the Turkish courts [108]. In countries with less transparency with their censorship, researchers and activists have conducted studies, tests, and measurements in an attempt to get a better understanding of the censorship mechanism of each country.

Aryan et al. [5] investigated Internet censorship by setting up a testbed in Iran to perform network measurements. They found Adult websites were the most blocked, and almost half of the top 500 websites of the Internet were also blocked in Iran.

Political activists and bloggers in Southeast Asia have been struggling with their freedom of speech. Vietnam imprisoned more than 46 bloggers and activists in the first half of 2013. Singapore's new rules governing online news led 150 Singaporean websites to blackout in protest in 2012. Thailand banned more than 20,000 URLs in 2012, causing a chilling effect on freedom of expression throughout the country [17]. Myanmar's military-led regime even limited communications by preventing access to the Internet and prohibited the use of communication technology equipment such as fax machines and satellite dishes [55].

Cubans have long been struggled with tightly controlled information for more than 50 years [29]. While Cubans have Internet access, it comes in the forms of email and local intranet where the only content available is the one hosted in Cuba. Moreover, Internet access is extremely expensive for Cubans [55]. Thus, email access is more common than Internet access because of the limited available content on the Internet [29, 55]. Shklovski and Kotamraju found similar results in an anonymized country where they conducted an interview study. Participants found blocking and censorship to be confusing and inspiring self-censorship. Participants also blamed

the lack of content on the nation's Internet and a threat to personal security on censorship. In the context of online contribution to sites that rely on user-generated content, Internet censorship creates conflicting goals between encouraging Internet users in the country to generate content and controlling certain types of speech on the Internet [95].

In the United States, Internet censorship is not as prominent, thanks to the free speech provision of the First Amendment of the US constitution which prohibits federal, state, and local governments from directly censoring the Internet with exceptions for obscenity, and especially child pornography. As a result, self-censorship by site operators is more common. Some institutions and web sites employ censorship in the form of content moderation to prevent controversy and inappropriate materials emerging from user-generated content. For example, several popular social media sites such as Facebook [33], Twitter [109], and Instagram [51] have explicitly (in their terms of service) reserved the right to remove inappropriate content. Nevertheless, these US-based social media companies still have to comply with government regulations in other markets they serve to ensure that the content served on their sites are deemed appropriate by other governments [61, 108].

2.2.2 Internet censorship practices in China

When the Internet was first introduced in China in 1990s, international observers suggested that the Internet technology would pose a threat to China's authoritarian regime [55]. The Chinese government has carefully controlled the Internet development in the country and developed one of the most notorious and the most technologically advanced [126] mechanism of Internet censorship: the Great Firewall of China (GFC). Because only a limited number of companies are licensed by the government to provide international network access to regional ISPs [13], the Chinese government can easily regulate international Internet traffic through GFC. GFC is known

for its strict and dynamic censorship patterns where all traffics going in and out of the country are inspected. Requests that match blocked keywords are restricted by the routers that regulate the country's Internet traffic [19]. In addition to specific keywords, popular social media sites such as Facebook and Twitter are also blocked in China [116]. However, the Chinese replicas of social network sites including Renren (a replica of Facebook) and Sina Weibo (a replica of Twitter) are allowed in China, albeit heavy content monitoring and censorship [60].

As more Chinese citizens become educated in the use of the Internet, they are more aware of foreign products, cultures, and norms. Internet users turn to social media to criticize the government and callout wrongdoings by governments and military officials [119], launching a "blog revolution" [31, 84] which thriving the Internet with creative usages of political satires, codewords, visual files, and implicit criticism [40, 47, 67, 69, 122, 125]. The revolution has turned the Internet into the platform for political debate.

The Chinese government thus felt the need to impose tighter censorship and put more responsibility of censorship on service providers [55, 100]. In physical space, Internet cafe owners are responsible for monitoring their patrons' digital activities [100, 121]. In digital space, chat room administrators need to hire censors or "Big Mamas" to screen and remove offensive materials from online bulletin boards [103]. With the rise of popularity of social networking sites, site operators still follow the same principles as in the early days of the Internet. On Sina Weibo, a Chinese microblogging service and one of the largest Chinese social media, there exists a set of terms that led to a higher rate of post deletion. Moreover, posts from conflicting regions were also deleted at a higher rate than those from other regions in China [8]. King et al. added that posts that promote collective actions—regardless of their pro- or anti-government point of view—are mainly censored [60].

Taking a step further, the Chinese government has recently "partnered" with technology companies such as China Mobile (a telecommunication provider,) Alibaba (an e-commerce site,) Tencent (an instant messaging service,) and Qihoo 360 (a security company providing anti-virus software, web browser, and mobile application store) to station police officers at these companies. While the companies claim that the police officers are there to "combat illegal and criminal activities on the Internet," many believe that this is a political censorship move by the government [24]. So far, the government has arrested more than 15,000 for "crimes that jeopardized [the Internet] security" [104]. On a positive note, a majority ($> 95\%$) of participants from a recent survey of Chinese Internet users were aware the existence of Chinese Internet censorship [112].

With the aim to fully explore the censorship apparatus behind Chinese social media, King and colleagues reverse-engineered the mechanics of censorship on Sina Weibo [61]. To do this, they set up a new social media company in China in order to gain access to customer service agents who would supply details about Chinese social media censorship. In addition to automated review through keyword matching, they found that a massive number of human censors also take part in the process. Figure 2 is a reproduction of their major result, representing a decision tree of censorship on Chinese social media.

Figure 2 explains how the automated censorship mechanism and human censors work together to filter inappropriate content on Chinese social media. Once a user submits a post to Chinese social media, the automated censorship mechanism detects whether the post contains any inappropriateness. If the automated mechanism clears the post, the post gets publishes to the site. However, human censors can still reevaluate the post at a later time, within 24 hours, and remove the post from the site if necessary. If the automated mechanism flags the post, human censors will manually review the post and decide to allow the post to be published or to delete to post.

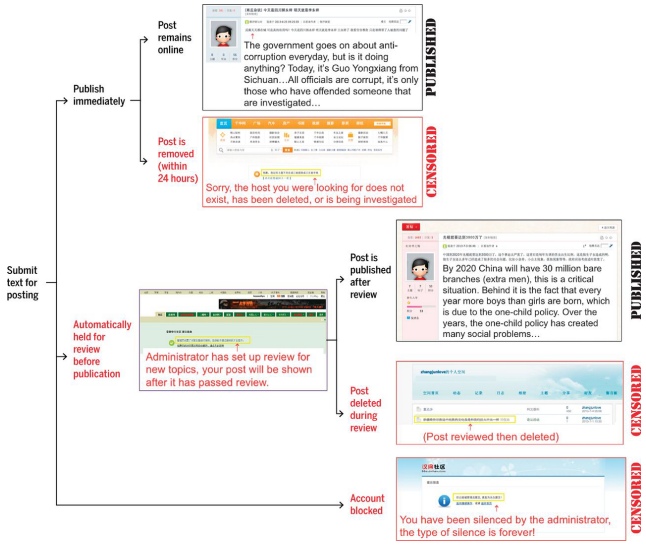


Figure 2: Chinese censorship decision tree, reproduced from King et al. [61].

If a user repeats posting inappropriate content, the account can get banned by the automated mechanism.

In my research, I plan to extend the current understandings of censorship practices and their consequences, especially in the context of Chinese social media. While previous research has extensively explored the mechanisms behind censorship on Chinese Internet and social media, little work has been done to explore the users' side of this equation. In this thesis, I conduct research to understand the effects censorship has on Chinese social media users, using both quantitative and qualitative methods in order to explore the breath and depth of this issue.

2.3 Social Media Use under Repressive Regimes

Social networking sites (SNS) bootstrap from real-world social connections and allow users to expand connections beyond their real-world friends, families, and acquaintances [11]. Researchers have been arguing that recreational use of media, whether traditional media such as televisions [81, 82] or modern media such as Internet [64, 92, 93], have disengaged people from participating in political and civic activities. However, a contrasting example was shown that exposure and attention to

public affairs television program can enhance political participation [72, 73, 77].

While social network sites provide opportunities to connect with people all over the worlds, their users do not use these sites to make new connections, but rather *bonding* existing ones [11, 65]. Social interactions on SNS can increase social capital even though the tie between the two persons is only a weak tie [14]. While the online interactions on SNS seem virtual and insubstantial, close friends and strong ties are still much more influential than weak ties [10]. However, only *bridging* social connections—connecting heterogeneous groups of people—will bring about social and political change [83]. Before 2008, there were no substantial evidence that show social media's influence on political activities [123]. However, the 2008 Obama Presidential Campaign has shown how social networking sites can play a critical role in social mobilization for political goals [20, 21, 90, 119].

Under repressive regimes, social networking sites are more than just places where people go to make connections. Because traditional media outlets usually get monitored and censored by governing units and journalists frequently practice self-censorship [99], political activists and citizen journalists turn to social media to broadcast their messages and report events in their local areas. Social media provides new sources of information the regimes cannot easily control [107] because censorship and blocking of the Internet can cause uproar against the governing units from not only political activists but also average citizens. The examples in Egypt and Tunisia have shown that technology can help overthrew repressive governments without the need of organized leaders [46].

The most notable case which social media played a big role in overthrowing repressive regimes was the 2011 Egyptian revolution. Egyptian dissidents and activists used social media platforms such as email, blogs, and social networking sites to create uproar and arrange protests against the regime of, then, President Mubarak. The organization of protests on social media was successful in gathering people in Tahir

Square, causing the government to shut down the Internet just after midnight on January 28, 2011. However, the shut down came too late and that morning, there were people at Tahir Square more than ever [28]. Al-Ani et al. conducted an analysis on Egyptian blogs and the roles they play in the 2011 Egyptian revolution [4]. Topic modeling analysis on the corpus showed a strong inverse relationship between the occurrence of personal/self-oriented posts and political posts. Overtime, Egyptian blogs became increasingly political. In addition to providing commentary regarding political situations, many blog posts reported updates on events occurring in Egypt during the Arab Spring. Several aspects of the uprising such as police presence and government reactions were included in the reports.

The use of social media in Egypt leading to the revolution and overthrown of Mubarak's regime was influential to other countries. For example, people in Tunisia, using a similar model of communications through social media, arranged a demonstration in Tunis with a nationwide call for participation mainly via Facebook [120]. The uproar in Tunisia started when the government blocked access to Dailymotion, a popular video sharing site among French-speaking countries, when activists posted a video unfolding how the president abused the use of the Tunisian presidential aircraft for private trips. Tunisian citizens, who might not be politically active, were made aware of the government's concern of free speech from this censorship [125].

The use of social media to arrange political protests can be traced back for more than a decade, predating the popularization of social networking sites such as Facebook and Twitter. One of the early protests organized with the aid of communication technology was in 2001 when citizens in Manila arranged a protest via text messages, resulting in a mass of a million people in downtown Manila [94]. Over the past decade, there were many instances where social media has played an important role in successful uprisings of citizens against dictatorship and repressive governments

such as Egypt [4, 28] and Tunisia [120]. Moreover, these uprisings also created ripple effects to other repressive governments who are trying to adapt to the evolution of technology to prohibit technology-led political movements [46, 94]. For example, the Chinese government blocked the search term "Egypt", fearing that the Egyptian protest would inspire unrest in China [7]. Research suggests that the use of social networking tools increases interpersonal discussion that fosters civic participation and political activism [7, 123], contrasting the early theories developed by social scientists [64, 81, 82, 92, 93].

These phenomenons played a part in inspiring the development of the "Cute Cat Theory of Digital Activism" [115, 125]. The theory posits that most people only use the Internet for mundane activities such as searching for pornography and images of cats. Tools such as Facebook, Flickr, Twitter, are developed for people to share this kind of contents with each other. These platforms are also useful to social movement and political activists who choose not to develop dedicated tools themselves. In turn, activists are more immune to government blocking and censorship because shutting down popular websites would provoke a larger public uproar than shutting down dedicated platforms for activism [115].

However, using social media does not always guarantee a successful demolition of repressive regimes. There were several cases that the activists' uprisings organized through social media have failed. For example, street protests in Belarus in March 2006 organized via email leaving the president "more determined than ever to control social media." The June 2009 uprising of Green Movement in Iran where activists exerted technological tools to organize the protest and the 2010 Red Shirt uprising in Thailand where social media savvy occupied downtown Bangkok both resulted in violence crackdowns [94]. HCI researchers started to focus on using technology to promote peace and reduce conflicts that could lead to war. The impact of war not only affects the countries' economic status [101] but also incurs costs in terms of

valuable human lives [45].

Using social media to organize gathering of citizens and political activists have surprised many researchers and journalists by the fact that the Internet can generate such strong commitments from its users. Of course, social media by itself cannot cause changes and revolutions. However, social media "combined with the right economic, social and political forces can be a potent threat to any leader, anywhere" [96]. Social media helps form social capital among people with the same struggle and same vision to create a bigger group to fight against repressive regimes. Wellman et al. raised a question how the Internet impacts social capital in real-world communities [114]. Increasing social capital, the Internet provides new and better ways of communication and meeting spaces for people with common interests, ridding of the limitation of space and time. On the other hand, the Internet could remove users from their immediate physical environment, and thus, reducing the social capital in the real world. In mediation, the Internet may be better at bonding existing social connections than creating new ones [63]. Thus, it is hard to generate organizational and political participation if users have no existing interests in the matters [114]. However, social media can function as the first step of engagement, along the line with the foot-in-the-door strategy which is a phenomenon when a person is more likely to fulfill a large request when he/she already agreed to a more modest request [37, 57].

Social media use has become essential in everyday lives of Chinese Internet users. However, there is a gap in research between the practice of censorship and user behavior in the context of Chinese social media. Little research has explored how censorship causes changes in the usage behavior of Chinese social media users. As a part of this thesis, I analyze usage logs of social media users who have been censored to find out the answers to the question of how censorship affects social media usage behavior.

2.4 Censorship Circumvention

There are two dimensions of anti-censorship tools: free access to information and free publication of information [66]. In this section, I review related work that documents and develops tools to approach each of these dimensions.

2.4.1 Free access to information

There are several strategies that people living in the countries with Internet censorship employ to get around censorship and retrieve information censored in their countries. Internet users with moderate to advanced technical skills rely on services such as VPN, proxy, and anonymizer to get access to blocked content and keep themselves anonymous in the case that their generated content cause any troubles [5, 95]. Services such as VPN Gate [110] provide access to VPN servers worldwide free of charge, allowing citizens of countries under censorship to access VPN services at little to no cost. Some Internet users utilize their social connections outside of their countries to gain access to blocked sites and to get around speed throttling [29, 95, 120].

VPN and proxies are proved to be potent in providing access to information censored. Thus, researchers have devoted efforts to create tools to measure censorship from single or several vantage points [16, 36, 49, 54, 78, 91]. When the Internet services are not available, wireless mesh network [1, 2] supports hyper-local networks where users can create their own networks. While the range of wireless mesh network does not span as large of an area as the Internet does, several situations have proven that this technology can be useful when the Internet services are congested [3], unavailable [53, 56], or inappropriate for the scope of communication [58, 71]. Cities in Greece and Spain have shown that wireless mesh network can span the area of a city with thousands of nodes in the network, threatening internet service providers to provide better services [6, 28, 41, 62, 105].

2.4.2 Free publication of information

With the Web 2.0, users become content generator rather than just content consumers [119]. However, not all regimes allow their citizens to speak freely on the Internet. I have already shown in previous sections that Chinese social media companies are required to hire human censors to filter content on their sites [61, 103]. Chinese social media is already filled with word plays, morphs, and homophones to circumvent censorship [18, 47, 122, 125]. Social media users often practice self-censorship to minimize the possibility of getting blocked [95].

Numerous researchers have worked on techniques and tools with the aim to circumvent censorship on the Internet. Similar to Chinese word plays, users can translate a message in to another language and back to the original language [86] or omit certain parts of a message [39] to create confusions to censors and become immune to surveillance. Researchers have also developed tools to encode hidden messages into regular, innocuous media [15, 35]. However, the pitfalls of these tools are (1) receivers might not share the same knowledge as senders to recover missing information or (2) messages need to be decoded using technologically advanced tools. Moreover, these tools are appropriate for only private communications, rendering useless when users wish to publicly broadcast messages.

Because Chinese users already have access to information on Chinese social media and access to foreign information can be obtained through VPN, the free access of information is not in the scope of this thesis. To achieve the goal of free publication of information, I extend current knowledge about the mechanism behind censorship on Chinese social media and develop a system that helps Chinese social media users publish messages that are costly to censor.

CHAPTER III

USER-LEVEL EFFECTS OF CENSORSHIP ON CHINESE SOCIAL MEDIA

> "The Chinese government just released a new law. If you post something [on Chinese social media] that's not true but it has been shared for 500 times, you will be responsible for that. What action they will take I don't know. Maybe they will be fined 5,000 yuan." — P6, illustrating the extent to which China can control the Internet within its borders.

China has arguably the world's most advanced Internet filtering system—the Great Firewall of China. It not only blocks specific sites, but also inspects every packet of Internet traffic to filter banned keywords [19]. Since people cannot access sites like Facebook and Twitter, Chinese social media services such as Sina Weibo and WeChat have flourished over the past decade [8]. However, unlike their Western counterparts, the operators of sites like Sina Weibo are required by the government to heavily censor inappropriate content using both advanced algorithms and human censors [61, 103]. From the state's point of view, this setup is ideal, as it allows the citizenry to have access to modern communication technologies as well as let off steam about governmental injustice [42], yet those technologies live under the control and surveillance of central authorities [70]. Previous research has explored the mechanisms behind the censorship [8, 61] as well as techniques that Chinese social media users employ to circumvent it [18, 47, 122, 125]. However, little is known about the effects censorship has on actual users of these censored systems.

This chapter presents a mixed-methods study focusing on user-level effects of censorship on Chinese social media. Borrowing from distributed cognition [48], the

effects were categorized into two types: *off-platform* effects and *on-platform* effects. Off-platform effects live inside users' minds and habits (i.e., self-censorship practices, risks associated with posting censored content) and are *not* directly observable in a user's social media usage log. On-platform effects are the opposite: they show in a user's trail of social media usage (i.e., effects on subsequent speech and account abandonment following the enactment of censorship).

The mixed-methods study involves two parts: a quantitative data analysis and an interview study. I gathered the data of more than 1.6 million Sina Weibo users, one of the largest Chinese social media sites—8,140 of whom were the subjects of government censorship. Using a propensity score matching design, I examine the on-platform effects of censorship by comparing censored users and a constructed control group. I also interviewed 11 Chinese social media users to learn the habits and behaviors users have adopted in response to widespread censorship.

A strong off-platform effect of censorship on Chinese social media was detected. Users heavily self-censor around political topics because of uncertain, perceived risks of censorship, echoing the findings from earlier work [95]. The matched-sample analysis illustrated on-platform effects: censored users reduced their posting activity 3.91% more than the control group in the 30-day period following censorship. That is, the enactment of censorship leads to a short-term suppression of speech. Moreover, 3.55% of censored users presumably abandoned their accounts in the same period, a small increase over the 1.33% of the control group. While the on-platform effects are present, they diminish over time, and eventually, as indicated by [112], users do not perceive the effects of censorship in their daily social media routines.

For those opposed to state-mandated censorship of social media, these results are discouraging. While both off-platform and on-platform effects certainly exist, the on-platform effects are relatively small and the off-platform effects are largely confined to controversial topics. In other words, it seems to us that the state is getting precisely

what it wants from its censorship apparatus: people think twice about posting about political topics, and state controlled social media platforms like Sina Weibo do not pay a steep price in user participation when it actually deploys censorship. This is crucial to their place as a "safety valve." Looking beyond China, it therefore seems difficult to argue—in terms of participation costs—that other authoritarian regimes (e.g, Russia, Iran, etc.) should not simply copy the Chinese model.

3.1 Research Questions

As outlined in the previous chapter the extent that the Chinese government has controlled their Internet, Internet users in China face with not only rigorous website and keyword filtering but also extensive machine and human censors on social media sites hosted in China. Ongoing work has attempted to document censorship practices in China [19, 23, 60, 61, 112]. However, most of them are conducted do not involve Chinese Internet users. Therefore, there is a gap in the exploration of the behaviors of Chinese Internet and social media users in the face of widespread censorship, leading to my first research question:

> **RQ1. How are Chinese social media users affected by censorship *off platform*?**

To get into more details, the literature suggested that censorship is heavily correlated to the content of posts, I want to further explore users' perception of censored content: [60, 122, 124].

> **RQ1.1. What content do Chinese social media users perceive to be censorship-prone?**

Moreover, Western media have reported several cases where activities on Chinese social media have led to prosecutions or disappearances of social media users. I want to assess the risks perceived by Chinese social media users: [9, 76, 79].

RQ1.2. What are Chinese social media users' perceived *online* and *real-life* risks of censorship?

Previous research has shown that Chinese citizens have mixed opinions on censorship [112]. Although [95] has shown self-reported data that censorship discourages contribution to sites with user-generated content, Chinese social media still adds users year after year. In 2015, an estimated 481M people (35.4% of the total Chinese population) reportedly visited social media at least once a month [50]. This mismatch leads to the next research question:

RQ2. What are the *on-platform* effects censorship has on Chinese social media users?

One of the key metrics for engagement on social media is user participation. As demonstrated by [95], although social media sites can be up and running, users may neglect the platform due to censorship.

RQ2.1. How does an act of censorship affect subsequent user participation?

Even though GFC blocks access to Western social media such as Facebook and Twitter, an increasing number of of Chinese Internet users (166M in 2014, 188M in 2015 [97, 98]) were reported to have access to VPNs and thus, have access to more selection of social media, including restricted ones. This leads to the next part of the research question:

RQ2.2. How does censorship influence user abandonment of Chinese social media accounts?

To summarize, I ask two main research questions in this chapter. First, what are the *off-platform* effects of censorship that are not observable from usage logs. Second, what are the *on-platform* effects of censorship that are visible from usage logs.

1. How are Chinese social media users affected by censorship *off platform*?

2. What are the on-platform effects censorship has on Chinese social media users?

3.2 Methods

I categorize the user-level effects of censorship on Chinese social media using the language borrowed from Distributed Cognition, a framework that describes how cognition is distributed across users, objects, artifacts, and tools in an environment [48]. Thus, I explore the research questions in two ways: *off-platform* and *on-platform* effects. I seek to understand both the effects of censorship on users' mental models and habits (off-platform), as well as their actual practices in response to the enactment of censorship on-site (on-platform).

In this chapter, I am interested in exploring perception of censored content *(RQ1.1)* and perceived risks of getting censored *(RQ1.2)*, both of which are *off-platform* effects and do not usually translate to observable signals on Chinese social media. Therefore, I choose to conduct an interview study to answer these research questions. I also take a close look at on-platform effects. *On-platform* effects are observable via data from Chinese social media platforms—such as user activity levels, the content of posts, etc. In this chapter, I explore participation *(RQ2.1)* and account abandonment *(RQ2.2)*.

In this section, I detail the methods of the mixed-methods study—both the interview study and the quantitative data analysis. I start with the details of the interview study. Then, I describe the datasets behind the quantitative analysis. Finally, I lay out the statistical analysis of the datasets to answer my research questions.

3.2.1 Interview Study

Due to the sensitivity of the topic of Internet censorship and to ensure that participants are familiar with Chinese social media and culture, I carefully selected participants based on a number of criteria:

1. To ensure that participants are familiar with culture, politics, and the Internet in China, participants must be Chinese citizens who have lived in China for at least two years.

2. To mitigate risks and protect participant identity and security, participants must be in the US at the time of the interview, so the interviews can be conducted either in person or over domestic US phone calls. Because I do not know the surveillance capabilities of the Chinese government inside China, I err on the caution to remove the risks that the interview sessions could be under the surveillance of the Chinese government.

3. To ensure that participants are familiar with Chinese social media, participants must be users of Chinese social media.

While this induces a bias in the selection of participants, I believe that these requirements are justified given the sensitivity of the topic and the risks associated with the study. I did not opt for Internet-based channels (e.g., Skype calls or a survey) because of these risks. To make sure that participants' identities are fully protected, I requested that Georgia Tech Institutional Review Board (IRB) waive documentation of consent. That is, the participants did not have to sign a consent form, leaving no record of their identities. Moreover, I also requested to forego the collection of participants' information for the purpose of compensation. Therefore, no personally identifiable information was kept. Other identifying information about the participants, such as age, gender, location, years living in the US, social media handle, etc., were also not collected.

I recruited participants through several of Georgia Tech mailing lists, personal contacts, and snowballing from recruited participants. In the end, I enrolled 11 participants in the interview study. During the interviews, participants were asked about their general use of Chinese social media: what sites/applications they use,

the purposes for which they use Chinese social media, etc. Then, they were asked questions to answer all of the research questions: user perception of censorship and responses to censorship on Chinese social media.

Interviews were semi-structured and conducted in English. The interview questions are presented below. While the participants' first language was not English, they were fluent in English. The interview lasted 20-45 minutes, with an average of 33 minutes. Participants were compensated with $15 retail gift cards for their time. The interview sessions were audio recorded, with the consent of participants, and transcribed. Then, I conducted thematic analysis [12] and performed qualitative coding based on the research questions I have established.

3.2.1.1 Interview Questions

Background Questions

1. What are the Chinese social media sites that you use?
   (a) What kinds of users do you follow?
   (b) What kinds of content do you post on social media?
2. Comparing Chinese social media and Western social media (Facebook, Twitter) how do you use them differently?
   (a) How often do you visit each of the sites? Tell us why your usage between Chinese and Western social media are different.
   (b) How are your posts on Chinese and Western social media different, in terms of content and the number of posts?
   (c) Why do you use Chinese and Western social media differently?

Acknowledgment of censorship

3. How did you come to know of the existence of censorship on Chinese social media?

4. Do you think Western social media (Facebook, Twitter) observe similar censorship? What made you think so?

5. Tell us about your experience of getting censored.

   (a) If so, how did you know that you have been censored?

   (b) What were the contents of the posts that were censored?

   (c) How did you feel when you were censored?

   (d) What did you do after you learned that your posts got censored?

6. Tell us about your experience of noticing other people's posts getting censored.

   (a) How did you come to notice them?

   (b) What were the contents?

   (c) How did you feel when you saw other people's posts get censored? How did it impact your use of social media?

7. In your opinion, what types of posts generally get censored? Why do you think so?

Effects of Censorship

8. How does censorship make you feel?

9. If censorship on Chinese social media did not exist, how would you use the sites differently?

   (a) Would you post more to the sites?

   (b) Would you spend more time on the sites?

10. Tell me about posts that you withheld from posting because of censorship

    (a) What was the content?

    (b) Who was your target audience?

    (c) What did you end up doing with that post(s)? Did you publish it/them elsewhere? Where else did you publish it/them?

11. If you use both Chinese and Western social media, how do you use them differently?

    (a) Who are your friends/connections on each site? How are they differ?

    (b) What content do you post to each site? How do they differ?

12. Have you ever thought of leaving Chinese social media due to censorship? Why or why not?

13. If Western social media are readily available in China, will you leave the Chinese sites? Why or why not?

Perceived risks of censorship

14. What do you think could happen to your social media account if you post something that gets censored?

15. Have you ever noticed something different with people you follow when they get censored? What was it?

16. What do you think are the risks of posting censored-sensitive posts on Chinese social media?

17. Tell us what do you think could happen if you keep posting posts that get censored?

(a) Why do you think so?

(b) Where/Who could these consequences be from?

18. What are the measures that you practice to stay safe on social media?

3.2.2 Sina Weibo Datasets

In parallel, I collected large-scale data from Sina Weibo, one of the largest Chinese social media platforms, with more than 200 million monthly active users [111]. Sina Weibo is a microblog service where users can post a short 140-character status update to their timeline, essentially the Chinese equivalent of Twitter. Because of its large user base, Sina Weibo has amassed a variety of users and content, ranging from daily updates from everyday people to political comments from journalists and activists.

I collected two datasets. First, censored posts were collected from Weiboscope [124], a site that curates Sina Weibo posts from popular accounts and also periodically checks whether these posts have been censored. I collected 42,638 posts that were censored from January 1, 2014 to October 3, 2015. These posts were authored by 9,860 different Sina Weibo users. However, as of January 1, 2016, 1,720 user accounts from this group were completely deleted. I discarded these users from our dataset because I cannot obtain their account information and the reasons these accounts were deleted: voluntarily or banned. Thus, 8,140 users were left in the first dataset.

For each of these users, I collected their basic profile information as of January 1, 2016, including gender, the number of followers, the number of posts, the date of the first post, and location. Moreover, I also collected recent posts from each of these users' timeline as of January 1, 2016. These users were assigned to be in the *treatment* group, as they had each been censored in the past.

The second dataset is the induced *control* group. I monitored the Sina Weibo public timeline and collected information including the number of followers, the number of posts, the date of the first post, and location from more than 1.6 million users. After

Table 1: Summary of the dataset. The leftmost column shows descriptive statistics of the control group before matching. The two rightmost columns (in green) show descriptive statistics of the matched control group and the treatment group, both of which I use in the quantitative analysis.

	Control		**Matched Control**		**Treatment**	
	mean	sd	mean	sd	mean	sd
Attributes						
number of followers	4,914	172,075	379,355	2,073,549	477,626	3,148,631
number of accounts follow	445	619	802	831	959	859
number of posts	3,393	5,567	16,024	31,526	17,896	40,519
account age (days)	1,187	525	1,489	539	1,494	551
number of censored posts					4.31	10.89
N	1,665,487		8,140		8,140	
Matching Covariates						
log(number of followers)	5.30	1.58	10.01	2.58	10.02	2.61
log(number of posts)	7.20	1.51	8.47	1.88	8.46	1.97
log(account age)	6.90	0.77	7.20	0.53	7.20	0.54
gender male	748,931		5,676		5,676	
female	916,556		2,464		2,464	

Mahalanobis distance matching with the treatment group (details below) the control group was narrowed down to 8,140 users, and these users' recent posts as of January 1, 2016 were collected the same way as the users in the treatment group. Table 1 summarizes descriptive statistics of the treatment group and the control group.

3.2.2.1 Mahalanobis Distance Matching

In order to account for the fact that user accounts in our treatment group may not conform to the general demographics of Sina Weibo and general social media users, I need to construct a subset of the control group which most resemble the treatment group [89]. I performed Mahalanobis Distance Matching (MDM) by using the R MatchIt package [44]. MDM is similar to Propensity Score Matching (PSM) as both of them are statistical techniques for data preprocessing, suitable in the case of causal

Table 2: Statistic tests of the matching covariates and their corresponding raw values between the matched control and treatment groups.

	Matched Groups				**Statistics**	
	Control		Treatment			
	mean	sd	mean	sd		
					Student t-test	
Matching Covariates					t	p
log number of followers	10.01	2.58	10.02	2.61	-0.34	0.74
log number of posts	8.47	1.88	8.46	1.97	0.24	0.81
log account age	7.20	0.53	7.20	0.54	0.1	0.92
					Mann-Whitney Test	
Attributes					U	p
number of followers	379,355	2,073,549	477,626	3,148,631	33065000	0.82
number of posts	16,024	31,526	17,896	40,519	33057000	0.81
account age (days)	1,489	539	1,494	551	32857000	0.36
N	8,140		8,140			

inference where the treatment group is exposed to the treatment condition, but no systematic methods are available to obtain a control group [26]. However, MDM calculates euclidean distance between samples, while PSM assigns a score to each sample based on the logistic regression model of the control samples. Thus, MDM matching better utilizes the information of the matching variables and their relative importance [59]. In my experiments, MDM provides better matching between my treatment and control groups, resulting in similar matched control samples to my treatment samples.

My matching paired treatment and control users that were similar in four observable characteristics: the log-scale number of followers, the log-scale number of posts, the log-scale of account age, and gender. The reason I chose to use log-scale of continuous variables (number of followers, number of posts, and account age) rather than the raw values was because the range of these values in the dataset is extremely

wide. It is a common method to employ in social media analysis. I used the nearest neighbor method with a 1:1 ratio: the algorithm matched one treatment user with the most similar control user, one at a time, until all the treatment users were matched with the same number of control users [44].

After performing propensity score matching, 8,140 users from the control group were matched with 8,140 users from the treatment group. Table 2 summarizes the statistical tests to compare the matching covariates and their corresponding attributes in raw scale (rather than log scale) between the matched control and treatment groups. I compared the three continuous matching covariates (log number of followers, log number of posts, and log account age) using Student t-test. The assumptions of the t-test are met in this case:

1. The sample sizes are sufficiently large ($N = 8,140$ in both groups).
2. Figure 3 shows that all three attributes normal distribution.
3. Homogeneity of variance is satisfied in the case of log number of followers and log account age.
4. The test is adjusted for the unequal variances of log number of posts.

The raw-value attributes (number of followers, number of posts, days of account age) are compared using the non-parametric Mann-Whitney test as they all do not meet the assumptions of Student t-test. We can see from the statistics tests that the Mahalanobis matching did a good job matching the control group with the treatment group as all the tests show no significant differences in all comparisons between the two matched groups.

3.2.3 Post-censorship Participation & Abandonment Analysis

To observe the on-platform effects of censorship, I look at two metrics: posting activity and account abandonment. First, I examined posting activities in the 30-day

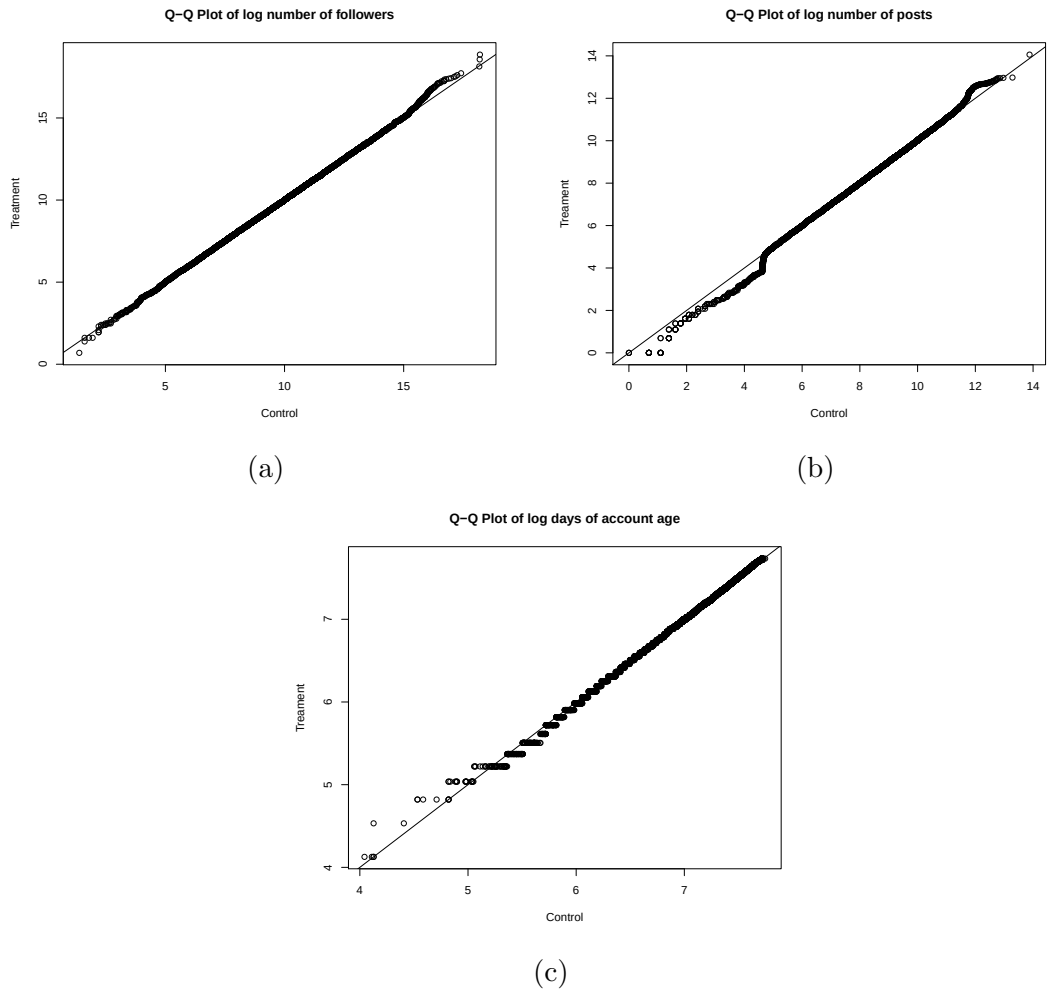


Figure 3: Q-Q Plots of three matching covariates between the control group and the treatment group. (a) log number of followers, (b) log number of posts, (c) log days of account age.

period before and after censorship occurred to explore the effects of censorship on user participation *(RQ2.1)*. To avoid an averaging effect in treatment users with multiple censorship instances, I only included the latest censorship instance from each user in the treatment group.

For each user in the control group, I randomly select one focus day from January 1, 2014 – October 3, 2015 (the days that users in the treatment group had their posts censored) on which the user had at least one post. Then, I gathered the posting activities from each of the users around each of their randomized focus dates. In

other words, the focus date serves as an artificial censored date for the control group. From hereon, I collectively call the dates of the latest censorship instances from the treatment group and the randomly selected dates from the control group the *focus dates*. As a result, each user in both the treatment group and the control group only contributed one interval around their respective focus dates.

Then, I look at how the effects of censorship on posting activity propagates through time. I varied the interval period of posting activity to 1, 3, 5, 7, 14, and 30 days before and after the focus dates of each user. Figure 4 shows examples of how the posting activity were calculated for different intervals. For each of these intervals, I compared the group total number of posts in the days before and after the focus dates between the treatment and the control group. Using Mann-Whitney U test, I test whether censorship in the treatment group has significant effects on posting activity compared to the control group.

To answer *RQ2.2* regarding user abandonment of Chinese social media account, I define *abandoned* user accounts in the k-day interval as the accounts that have posting activities in the k days before the focused dates, but not in the k days after the focused dates, $k \in \{1, 3, 5, 7, 14, 30\}$. For example, if Jackie's account has posts on Day -6 and Day 17 in addition to the focus date (6 days before the focus date and 17 days after the focus date). Her account will be considered as *abandoned* in the $k \in \{7, 14\}$ interval because of the activity in the k days before the focused date, but no activity in the k days after. For other values of k, her account is not considered abandoned.

I observe how the abandonment of accounts in both the treatment group and the control group changes over periods of time to conclude whether censorship has an effect on account abandonment. Note that I use the term *abandonment* as a presumption since I do not have the ground truth whether the users have returned to their Sina Weibo accounts after the data collection.

Figure 4: An example timeline in the quantitative analysis. If a user's focus date is July 31, we will focus on the posting activity from July 1–August 30. The 7-day interval will span from July 24–August 7 (7 days before and 7 days after the focus date). The 30-day interval will span from July 1–August 30 (30 days before and 30 days after the focus date.)

3.3 Results

Next I report results from our mixed-methods study, categorized by off- and on-platform effects. Before I present the results, I first contextualize Chinese social media using quotes from our interview participants

3.3.1 Chinese Social Media Landscape

The most popular services used by my participants were WeChat and Sina Weibo. WeChat is a messaging/social networking service where users can send messages to friends in their contact lists (similar to Whatsapp and Facebook Messenger). In addition to messaging, WeChat also has a "Moments" page where users can privately share status updates to their friends and view activities shared by their friends or public accounts they follow. Sina Weibo is a microblog service and has the same functionality as Twitter. Users can publicly or privately post a short 140-character microblog, or *weibo*. Regardless of the privacy option, most of Sina Weibo users post

publicly, and all of the participants perceived Sina Weibo as a public platform. Some participants even considered Sina Weibo to be "too public."

Although Western social media such as Facebook and Twitter are blocked by GFC [8], access to these services within China is possible via VPNs. Among my participants, only P11 started using Facebook while he/she was in China. The rest of the participants have heard of Western social media services while they were in China, but they were not interested in signing up because of the lack of known contacts on the platform. It was not until our participants came to the US that they started signing up for Western social media accounts.

All but one of our participants still use Chinese social media while they are in the US to keep contact with their friends and families in China, and to read up on news, current events, and trends. P5 was the only participant who reported that he/she no longer used Chinese social media because of his/her frustration with its censorship policy. His/her family had all migrated to the US, leaving him/her with no significant ties in China, and consequently, no reason to maintain Chinese social media accounts.

3.3.1.1 Awareness of Censorship

In order for censorship on Chinese social media to have an effect on users, the users need to be aware of censorship first. Previous research has shown that majority of Chinese Internet users are aware of censorship on the Chinese Internet and social media [112]. My results echo this finding. All but one participant (P4) reported that they knew of the existence of censorship on Chinese social media. Most of the participants knew about censorship from their friends or family members, while some experienced censorship first hand—having either been censored themselves or seen other users' posts censored.

Censorship is a common knowledge among Chinese Internet users. Sometimes if you talk about political stuff, because we are a communist country, we also have some bad stuff happened in the history I think the government still tries not to publicly talk about. (P2)

For those who encountered censorship first hand, they utilized several signals to detect censorship, such as disappearance of posts, system messages, and complaints from censored users.

[I posted something on Sina Weibo, and] within a few hours, this post was retweeted over a thousand times, had over 300–400 comments, and then it disappeared. Some people were asking me did I delete it; I said no. So I went back to take a look. First, I found out on my post, the [button] to retweet was gone. Secondly, it became invisible to outsider... Nobody can find it, only I can see it. (P1)

One day I posted something that used the year 1989. It's actually nothing to do with any event but my article was deleted because of some sensitive words in it and I was surprised and I tested it paragraph by paragraph then I found out the sensitive word is the year. (P7)

He [whose post got censored] just posted status that the picture [he] just uploaded got censored. (P9)

After the participants had experienced Western social media when they came to the US, they noticed the differences in censorship policy between Chinese and Western social media.

I feel like in America the censorship is not focused on politics. It still exists. In America people focus more on terrorism than politics. (P2)

I'm sure Facebook has the same capability of monitoring as Chinese social media but they don't exercise that the way they do in China. (P5)

3.3.2 RQ1. Off-platform Effects

Next, I present findings from our interview study addressing RQ1.1 and RQ1.2.

3.3.2.1 RQ1.1. Participants perceive political content as more censorship-prone.

The majority of participants believed that posts with content related to Chinese politics, governmental policy, and *"different opinions about [the government]"* (P11) were more likely to be censored than other content. However, none of the participants were able to specify what specific topics related to politics are being censored on Chinese social media. Participants vaguely claimed that the degree of "sensitivity" of the content is the deciding factor, but they were not able to quantify it or give concrete examples.

On a related note, P1, P6, and P9 referred to the recent crackdown of rumors on Chinese social media. Chinese social media providers have installed a "rumor clarification system" to automatically block rumors from spreading on the sites. At the same time, the Chinese government has passed a law to criminalize social media users whose posted rumors got more than 500 reposts [9].

> *The Chinese government just released a new law. If you post something that's not true but it has been shared for 500 times, you will be responsible for that. What action they will take I don't know. Maybe they will be fined 5,000 yuan. (P6)*

> *Sina Weibo has a system called rumor clarification system. If someone posts a piece of news that is not verified but later being verified fake, they will be punished. (P9)*

Mirroring the findings from [112], three of the participants endorsed censorship on the platform. They felt that censorship on Chinese social media is an appropriate measure to control a country with a large population.

In China, we have so many people, and if someone says something stupid there will be a lot of people who don't know that's not true. You can spread out wrong information. It's hard to rule the country that has so many people, so they want to make sure that good stuff is online. (P2)

However, other participants felt that censorship had become less useful and more irritating to Chinese Internet users—especially among young people.

I feel [censorship is] less and less useful in the new world. . . . Nowadays many people know how to get around the blocking or censorship and people just know more about the world than before. It's less and less useful and more and more agitating to people in China, especially young people. I think I do too. I feel that it's a restriction. (P7)

3.3.2.2 RQ1.2. Unclear online and real-life risks from censorship.

The participants valued the connections they had on social media. Therefore, they needed to adapt their social media usage to keep their accounts from being banned.

Sina Weibo is a very good social media tool to convey my message . . . I enjoy it. I love the interaction. So I'm very careful not to cross the red line . . . I'm trying to post something and having a lot of fun interactions but I also appreciate the social media outlet . . . I tried my best to stay within the boundaries. (P1)

The first risk of posting sensitive content or false information on Chinese social media that the participants observed was getting the post deleted. However, the participants had conflicting reports regarding what could happen to their online accounts. A few believed there are no consequences to posting sensitive content. Others believed that posting sensitive content could lead to banning or temporary account blocking.

Beyond online punishment, the participants also had conflicting concepts of real-life consequences or prosecutions from posting sensitive content. Due to a large population, a few participants did not foresee being tracked down from their social media profile.

However, the majority of the participants believed that getting censored on Chinese social media could have real-life impact. Most participants agreed that if the censored posts contain extremely sensitive content, then the government might take action to prosecute the social media users. Nevertheless, none of the participants knew for sure what kinds of sensitive content would trigger the prosecution and what actions the government would take against users.

> *I've heard of police involvement with someone posting anything too extreme. I'm not sure exactly what the police do with about it. They do take it to another level. (P5)*
>
> *They will be fined by the government. I think whatever you post, you have to be responsible for that. (P6)*
>
> *I'd say if you really bad rumors that have a bad social impact, you may get arrested. (P9)*

3.3.2.3 Off-platform Effects: Participants avoid creating original content to stay safe.

Regardless of the participants' mental models of censorship on Chinese social media and their perceived risks of posting sensitive content on social media, all of the participants were aware that they should be careful when posting content on Chinese social media. Even though some participants did not foresee any risks associated with posting sensitive content on Chinese social media, they still stayed away from posting sensitive content because they did not want to get involved in heated discussions on social media. In other words, the participants' awareness of censorship influenced

them to self-censor.

> *I would probably withhold anything political, anything opinionated towards the government. When the Shanghai expo was happening, I think I mentioned something like it was not that important of an event and I thought twice about posting that so I didn't. I couldn't access the VPN so I didn't post that on Facebook either. I didn't post it on [Sina] Weibo because I think it will end up getting censored. (P5)*

> *I might [post more if there is no censorship]. In most people's mind they have this self-censorship thing not only about political stuff. You think about it before you post it. There two ways to think about this, if it's inappropriate materials, then I don't think I'll post more of them . . . For political discussion, I might post more. I might feel more free to discuss about these things if there's no censorship against any political stuff. (P10)*

My participants also avoided the "responsibility" of posting sensitive content by using the reposting functionality on Chinese social media to echo the controversial opinion they agreed with instead of authoring their own posts. The participants felt that they would not be held responsible for creating sensitive content on Chinese social media by reposting.

> *If I see something I agree with, I would repost that post instead of stating my own opinions because if I had posted my own opinions, I get troubles with that. If I repost something I agree with, I wouldn't be responsible if they try to prosecute me. (P11)*

3.3.3 RQ2. On-platform Effects

Next I examine the effect censorship has on-platform: the residual traces of behavior after an act of censorship occurs.

3.3.3.1 RQ2.1. Censorship suppresses speech, but the effect wears out over time.

For a few participants who had their posts censored on Sina Weibo, the treatment of censorship did not drastically change their social media behavior. Rather, censorship cultivated a sense of caution.

> *For that article [that was censored], I just deleted. I changed the way to say my things and deleted the sensitive word. After that I don't know what the sensitive words are so I was aware of those and more conscious than before but still it's a passive thing and I don't know until they remind me if it is a sensitive content or not. I don't remember if my behavior changed or not but at least in my mind, there's some change. (P7)*

> *I didn't do anything differently. It just like a demonstration of censorship on the Internet. Before my post was deleted by the government, I just heard of this kind of thing and I saw something like that but it happened on someone else. But when my post was deleted I just realized everything is true. The censorship really exists. I'm quite used to it. I'm really familiar with this thing but I think it's kinda useless. (P8)*

However, what happens when we look at scale? Next, I examine the total posting activities 30 days before and after the focus dates of our treatment and control groups. Figure 5 shows the chart of the total posting activities, by day, from both the treatment group and the control group. The y-axis shows the total number of posts per day from each group, i.e., the sum of posts per day per user. The x-axis shows the day away from the focus date and centered on Day 0, the focus date.

The plot shows spikes on Day 0 in both groups because users had to post at least once on their respective focus dates (the definition of focus dates). We can observe from this plot a similar pattern from users in both the treatment and control groups. Users in both groups increased in posting activities a few days before Day 0. However,

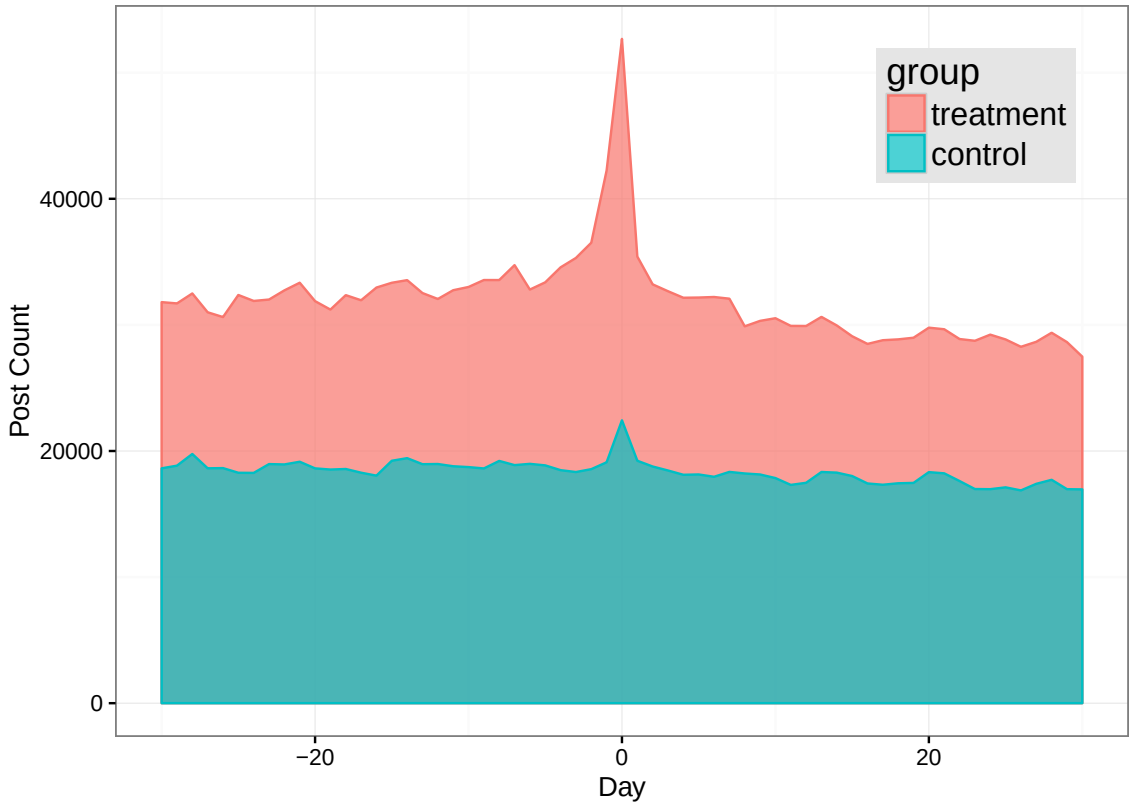


Figure 5: Comparison of the total posting activities 30 days before and after the focus dates between the control and the treatment groups. Spikes on Day 0 represent the definition of focus dates. See section for details.

on Day 0, users in the treatment group have a larger increase in the posting intensity than the control group—17.4% increase from Day -1 to Day 0 in the control group vs 24.7% increase in the treatment group. Then, the posting activities drop after Day 0. When we look closer around the center of the x-axis, we can see that the treatment-group activities started to steeply increase 2 days before Day 0. Then, the activities sharply drop right after Day 0.

To examine how the on-platform effect of censorship propagates through time, I compared the group posting activities before and after the focused dates in different intervals. Figure 6 shows the comparison of group posting activities from the treatment group and the control group in the intervals of 1, 3, 5, 7, 14, and 30 days before and after the focus dates.

Table 3: Proportion of change in group posting activities and paired Mann-Whitney test of posting activities before and after the focused dates, by interval.

Interval	Control		Treatment	
	change	U	change	U
1	0.64%	2233800	-16.15%	4779500***
3	0.81%	20800000	-11.19%	45257000***
5	-0.68%	53678000*	-9.00%	127030000***
7	-1.67%	105870000**	-7.87%	250550000***
14	-3.51%	432170000***	-8.21%	983250000***
30	-5.29%	2049600000***	-9.20%	4565400000***

* $p < 0.05$, ** $p < 0.001$, *** $p < 0.0001$

Figure 6 shows that the differences between the control-group posting activity before and after the focus dates do not differ significantly in most intervals. On the other hand, the treatment-group posting levels before and after censorship are significantly different in all intervals. For all intervals, speech is reduced.

To confirm the visual findings from Figure 6, I perform paired Mann-Whitney U tests on the posting activities before and after the focus dates to see if the changes in activity are statistically significant. Table 3 shows the results of those tests.

Paired Mann-Whitney tests confirm that in short intervals, the changes in control-group posting activity were not statistically significant. On the other hand, in the longer intervals—7, 14, and 30 days—the reductions in posting activity were statistically significant.

In contrast, the reductions in treatment-group posting activity are statistically significant in all intervals ($p < 10^{-16}$ in all intervals). The drop in posting activity is more drastic when in the intervals immediately after censorship than the farther out intervals. In comparison to the control group, having posts censored definitely has an effect on the treatment users: it significantly (although not dramatically) reduces their posting activity.

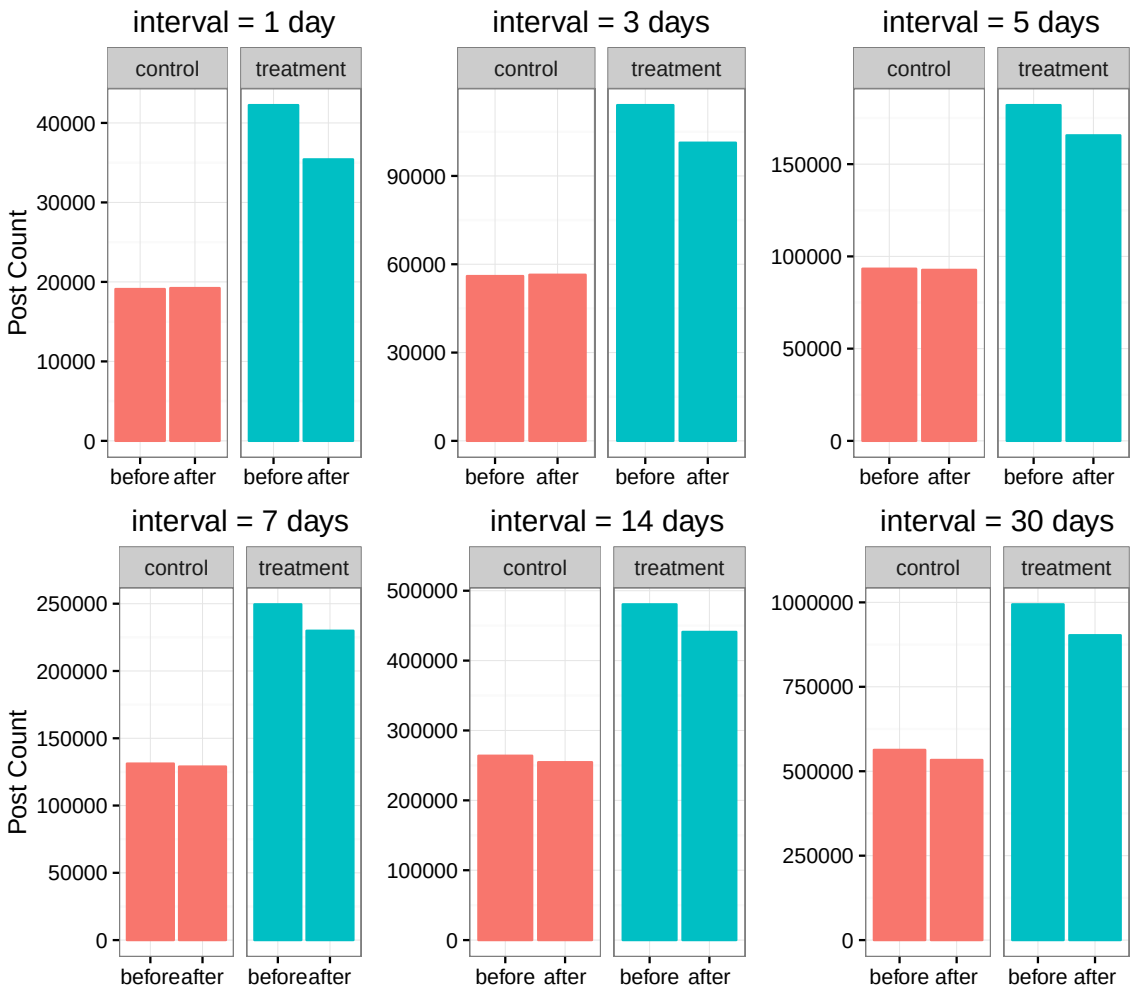


Figure 6: Group posting activities by interval around focus dates.

3.3.3.2 RQ2.2. Censorship does not, in any meaningful way, drive users away from Chinese social media.

Beyond posting activity, I also inspected the effect of censorship at the account level. In the same intervals as before, I totaled how many users have abandoned their accounts by group. Figure 7 shows the plot of proportion of users who abandoned their accounts in each interval by group.

The graph shows the proportion of users who abandoned their accounts are smaller once the interval is wider. Moreover, in all intervals, more users in the treatment

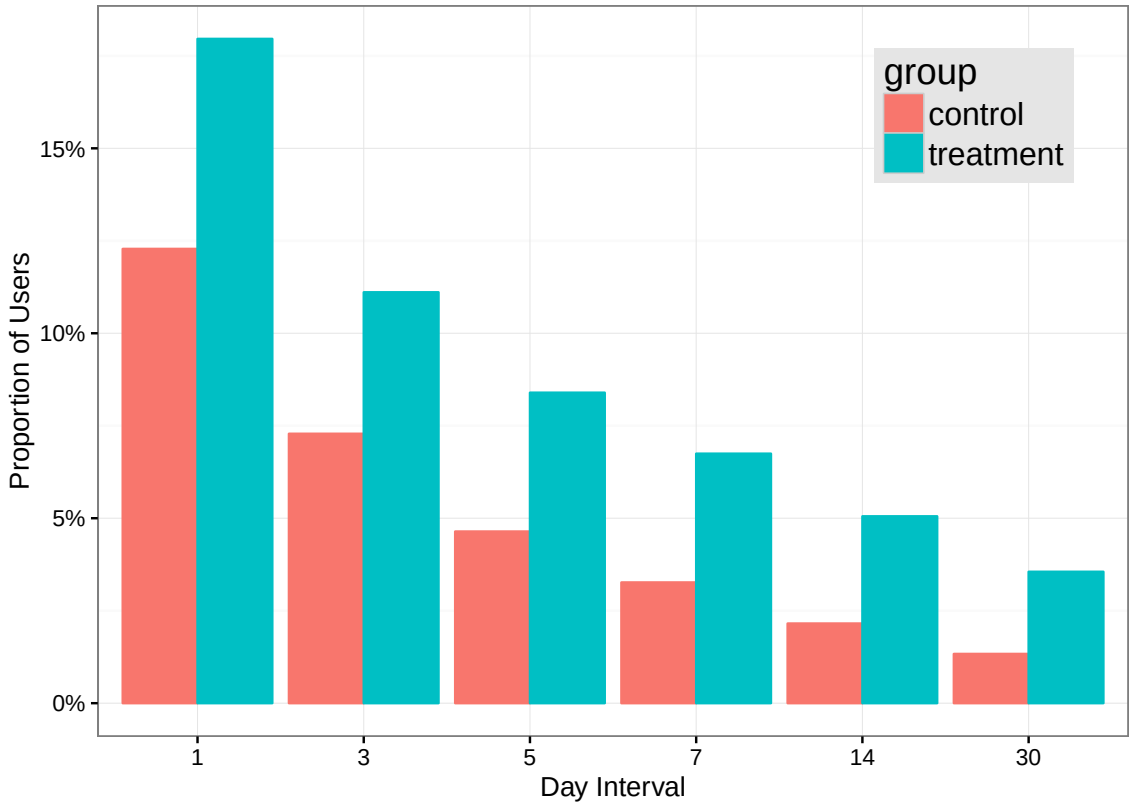


Figure 7: Proportion of users from the treatment and control group who abandoned their accounts in each interval.

group abandoned their accounts than the control group. Table 4 shows the proportion of users who abandoned their accounts and corresponding results of non-parametric equality of proportion tests without continuity correction between the treatment group and the control group, for each interval. It confirms the visual findings that the proportions of abandoned treatment user accounts are significantly different than the control.

However, in the 30-day interval, the abandoned proportion of the treatment group starts to catch up with the proportion of the control group. If this trend continues, we might see results showing that the abandoned proportion of both groups are equal once the interval period gets longer.

Our interview participants confirmed this finding from the quantitative analysis.

Table 4: Proportion of users from the treatment and control group who abandoned their accounts in each interval and their corresponding equality of proportions tests.

Interval	Control	Treatment	$\chi^2(df = 1)$	p
1	12.28%	17.96%	65.578	10^{-16}
3	7.28%	11.11%	45.432	10^{-11}
5	4.64%	8.39%	59.093	10^{-14}
7	3.26%	6.75%	64.418	10^{-15}
14	2.15%	5.05%	60.462	10^{-15}
30	1.33%	3.55%	51.220	10^{-13}

Despite the awareness of censorship on Chinese social media and its perceived effects, almost all of our participants agreed that they still want to keep using Chinese social media, even though they now have access to Western social media. Unsurprisingly, the main reason to keep using Chinese social media was to maintain contact with their friends and family members who are still in China and only have access to Chinese social media.

> *No, [I will] definitely not [stop using Chinese social media], because we just connect with friends and follow some celebrities. Most users of Western social media are westerners and they are not that close to Chinese people. I don't think I will quit WeChat or [Sina] Weibo even I can use Facebook or something else. (P8)*

3.3.3.3 On-platform Effects: Chinese social media users temporarily reduce their participation in response to censorship.

The findings show that in both RQ2.1 and RQ2.2, the on-platform effects of censorship are stronger in the short period following censorship. In RQ2.1, censored users reduced their posting activity more heavily immediately after censorship. Also, RQ2.2 shows that users eventually effectively return to Chinese social media regardless of their treatment of censorship. All the effects that we observed are temporary,

and the magnitude of the effects shrinks over time.

From a design perspective, Chinese social media has been bringing users into other activities besides posting status updates. A considerable portion of the participants claimed that they prefer to keep using Chinese social media over Western social media because the design of Chinese social media and its features were more suitable to Chinese culture. As P2 reported, there were more activities that users can perform on Chinese social media such as buying movie tickets and sending money to friends, none of which cannot be accomplished on Western social media.

3.4 Discussion

Borrowing the language of Distributed Cognition [48], we saw that censorship on Chinese social media has influenced user behavior more *off-platform* than *on-platform*. Chinese social media users did feel the effects of censorship but these effects were not translated into very large observable online behaviors. Next, I review and discuss each of the findings surrounding our research questions in turn.

3.4.1 RQ1. Off-platform Self-censorship Around Controversial Topics

Previous research [88] observed Chinese social media users rephrasing their posts once they get censored. Looking at RQ1.1 and RQ1.2 together, my study found that there is a mismatch between users' concept of censorship and the actual censorship mechanisms behind Chinese social media. This discrepancy leads to an inability to understand censorship signals, so users might try to make sense of the censorship model by conducting their own little experiments to better understand the censorship mechanisms.

On the contrary, once user cognition becomes too overloaded, they might just give up posting, as [95] reported that censorship discouraged social media users to contribute to the sites. Although the participants were aware of censorship on the platform, no one knew for sure how censorship works. Before posting to Chinese

social media, users were unable to make an informed decision whether their posts would be censored or not (RQ1.1). Consequently, users were also unclear about what risks—both online and in real life—were associated with getting censored on social media (RQ1.2). Therefore, self-censoring seems to be the way out for users, both to reduce the cognitive load and to avoid any unforeseeable consequences with their posting on Chinese social media.

3.4.2 RQ2. On-platform Effects Diminish Over Time

The findings from the quantitative analysis illustrate that censorship has statistically significant effects on user participation on Chinese social media. But those effects are small in real terms. In the 30-day period after censorship, the treatment group users lowered their participation on Chinese social media by 3.91% more than the control group. Additionally, the number of treatment users who presumably abandoned their accounts in the same time period was 2.22% more than the control group. However, as I presented in the previous section, these numbers become smaller in magnitude once the time intervals grow.

Our interview participants complemented the findings by expressing that they did not feel that censorship had an effect on their usage of Chinese social media. Furthermore, those who experienced censorship reported that the effects were *intrinsic* rather than extrinsic. In other words, censorship has taken root in the participants' minds; it has become automatic for our participants to self-censor before they get to the text box to post.

In this study, the k-day abandonment of censored Chinese social media accounts peaked immediately after censorship, but eventual account abandonment was rare. From the interview findings, participants report remaining on Chinese social media to keep contact with their friends and family members in China. Some even preferred Chinese social media services over their Western counterparts because of designs and

features that integrate into their daily routines.

3.5 Limitations

One of the major limitations in this work is my selection of interview participants. As addressed in the Methods section, I only included the interview participants who were physically in the US to limit the risks to participants. This decision heavily skewed the participant pool. Therefore, my interview participant pool was not a representative of general Chinese social media users. The interview responses seen in my study can potentially be one-sided due to the participants having higher education levels than average Chinese Internet users. Previous research suggests that Chinese Internet users with higher education levels tend to have an opposing view towards censorship on the Internet [112].

Nevertheless, my interview questions mainly targeted the participants' experiences with censorship on the Chinese Internet rather than their opinions and attitudes towards the practice of censorship. Moreover, the results presented show mixed responses from participants regarding their experiences with censorship. There is no perfect way to achieve a representative sample of Chinese Internet users: conducting this study within China could also lead to a skewed participant pool because of the self-selection of participants. I believe that the choices I made in this study were justified by ethical concerns.

The other limitation is with the datasets. While Sina Weibo does provide open APIs for developers to interact with the platform, the APIs are heavily restricted and offer little to no access to researchers for the purpose of gathering user information. I had to resort to the data collection from the web interface as done by previous researchers [43, 60]. Hence, I was not able to gather as much data as I would like to explore the long-term effects of censorship.

3.6 Design Implications

As a social computing researcher, I see an opportunity to help Chinese social media users in the process of authoring posts. As discussed in the previous section, there are many factors that are involved in users' decision making process when posting. One factor which heavily impacts user participation is the confusing and opaque censorship apparatus. Although there are numerous works in the literature that aim to help users circumvent censorship [15, 35, 39, 86], none of them integrates into users' social media routines. By using the technology of [124], for example, I see an opportunity to make censorship on Chinese social media more transparent to users. This way, users can have a clearer understanding of what content is censored at what time, as opposed to the current practice of the censorship "guessing game."

In the next chapters, I outline the algorithm that eventually drives the I system that will relieve the cognitive load of Chinese social media users by bringing more awareness of censorship to Chinese social media users.

CHAPTER IV

ALGORITHMICALLY BYPASSING CENSORSHIP ON CHINESE SOCIAL MEDIA

As with traditional media, social media in China exists under the watchful eyes of government censors. Censorship on Chinese Internet has been established since its inception. Without much relaxation from the government, Internet users in China gradually accept that censorship is *normal* and adapt to live with it [112]. Nevertheless, in limited cases, activists have employed *homophones* of censored keywords to avoid detection by keyword matching algorithms. Based on King et al.'s censorship decision tree [61], I speculated that it may be possible to consistently subvert censorship mechanisms by bypassing the initial review, thereby increasing the chance that posts will be published immediately. The key insight is to computationally alter the content of a post by replacing censored keywords with their homophones. As this is already an emergent practice on Chinese social media today [18, 38, 47, 125], I expected that the transformation would still allow native speakers to understand the original intent of the posts, given their awareness of the general topic of the posts. At the same time, the use of homophones may also allow the posts to bypass automatic keyword detection, since the posts no longer contain censored keywords. Ideally, the process of generating homophones to replace censored keywords would also not converge on only a handful of homophones for any given censored keyword. If it did, the censorship apparatus could easily augment their keyword dictionaries with commonly used homophones; rather, a non-deterministic, "maximum entropy" approach would likely add confusion and workload to current censorship apparatus.

In order to develop such algorithm, I chose Sina Weibo, the largest Chinese social

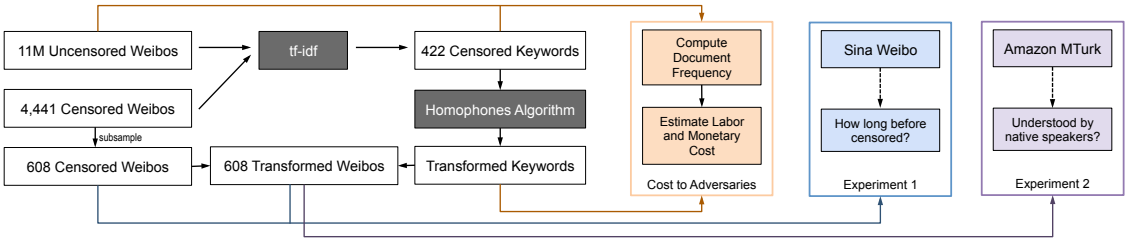


Figure 8: An overview of the datasets, methods, algorithms and experiments.

media with more than 222 million monthly active users [111], to be the target platform for this experiment. There were three research questions that I answered through this part of work:

RQ1. Are homophone-transformed posts treated differently from ones that would have otherwise been censored? Do they bypass the existing censorship apparatus, lasting longer on Sina Weibo?

RQ2. Are homophone-transformed posts understandable by native Chinese speakers? In transformed posts, can native speakers identify transformed terms and their original forms?

RQ3. If so, in what rational ways might Sina Weibo's censorship mechanisms respond? What costs may be associated with those adaptations?

Figure 8 presents an overview of the methods employed in this work.

4.1 Datasets

I obtained two datasets to explore these research questions. The first dataset consists of 4,441 posts that were confirmed to be censored on Sina Weibo. The dataset was gathered from the site Freeweibo[1]; Freeweibo curates posts from popular accounts on Sina Weibo and detects whether each one has been censored. Freeweibo also displays the top 10 "hot search" keywords that were searched through their website at any unspecified time period. I obtained all hot search keywords that contain only Chinese

[1] `https://freeweibo.com/en`

characters over a roughly one-month period from October 13, 2014–November 20, 2014, resulting in 43 keywords.

Because Freeweibo does not overtly indicate why each weibo was censored, I assume as ground truth that the hot search keywords were the factor that led to censorship. I believe that the hot search keywords are a good indication of censored keywords because of the high frequency for which they were searched on Freeweibo. If these keywords were not censored, people could simply do a search for them on Sina Weibo. In this manner, I collected a dataset of 4,441 censored weibos which were posted from October 2, 2009–November 20, 2014.

The second dataset consists of posts from the public timeline of Sina Weibo. I used the Sina Weibo Open API to obtain these weibos available, again from October 13, 2014–November 20,2014, accumulating 11,712,617 weibos.

4.2 Methods

4.2.1 Censored keyword extraction

Puns and morphs are only a few examples of how the usage of Chinese language in the context of social media often does not follow what is seen in dictionaries. Therefore, I decided against using a pre-existing dictionary to extract words and phrases from my censored weibo dataset. Instead, I generated all two, three, and four-character words/phrases from the censored weibo dataset. The terms that appear less than 10 times in the combined dataset of censored and uncensored weibos were then removed to ensure that the remaining terms commonly appear in social media. Then, I used the term frequency, inverse document frequency (*tf-idf*) algorithm to calculate the tf-idf score for each of these terms against the uncensored weibo dataset, treating each weibo as one document. I considered terms with tf-idf score in the top-decile to likely be censored keywords. I added to this computationally-inferred list the the hot search keywords from Freeweibo. In total, I therefore have 608 unique combinations

of censored keywords. For each combination, I took the latest weibo in the censored dataset to form the small dataset of 608 weibos to explore in my experiments. (My experimental methodologies, explained in greater detail later, carry a cost associated with each weibo in the dataset. I created a subsample for this reason.)

4.2.2 Homophone generation

Chinese words are a combination of several characters. Each character is a monosyllable and usually depicts its own meaning, contributing to the meaning of the larger word. Due to the racial and cultural diversity in China, there are numerous dialects of the spoken language, but only one standardized form of written scripts. In this work, I focus on Mandarin Chinese, China's official language. Mandarin Chinese is a tonal language: each character's sound can be decomposed to a root sound and its tone. Some characters convey multiple meanings and might be associated with multiple sounds based on the meanings they convey. While the tone of a sound can change a word's meaning, native speakers can often detect an incorrect tone by referring to its surrounding context.

Each Chinese character appears in written Chinese with a certain frequency—information my homophone generation procedure employs (to avoid generating very rare terms). I calculated the character frequency from my Sina Weibo public timeline corpus, consisting of 12,166 characters with 419 distinct root sounds (ignoring tones). There are 3,365 characters that have more than one root sound. For those characters, I assigned the frequency of the character to all sounds equally since I do not have information about the frequency distribution of the sounds. Then, for each of the 419 root sounds, I calculated the percentile of each character with that root sound based on its frequency from Da's character frequency list of Classical and Modern Chinese [25] to generate a frequency score for each Chinese character.

To summarize, for a character c with corresponding sound r, I calculated its

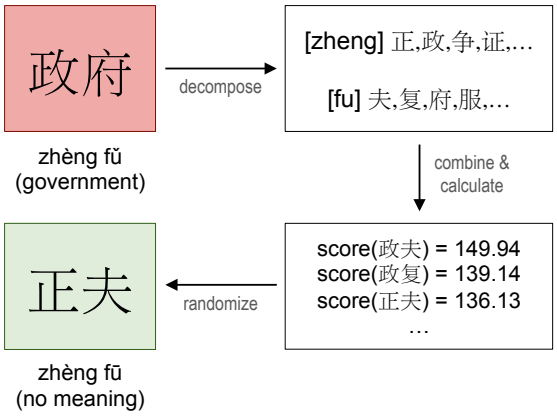


Figure 9: A high-level overview of the homophone generation algorithm.

percentile p based on its frequency compared to other characters that also have the sound r. For each censored word W with characters $w^1w^2 \dots w^n$, its homophones $\widetilde{W_i}$ were obtained by combining the homophones of each character $\tilde{w}_i^1\tilde{w}_i^2 \dots \tilde{w}_i^n$. Then, I used the following heuristic to calculate a frequency score for a homophone:

$$score(\widetilde{W_i}) = \sum_{k=1}^{n} p(\tilde{w}_i^k)$$

where p is the function that returns the sound percentile of its character parameter. Figure 9 shows an example of my algorithm generating a homophone for the censored keyword 政府(government).

Because the characters in the public timeline corpus might include archaic and rarely used characters, I picked the homophones $\widetilde{W_i}$ that have scores among the top k to penalize ones that include characters that might be unfamiliar to native speakers (low frequency). To ensure that the algorithm doesn't converge on the same homophone every time, one homophone out of the top k was randomly selected each time a homophone is requested for W. (In my experiments, I let $k = 20$.) Note that the algorithm has a high chance to generate homophones that have no meaning since I did not consult a dictionary.

Because the algorithm ultimately interacts with censorship adversaries (something

I describe in more detail in the *Cost to adversaries* section), I chose to shorten homophones of long censored keywords (4 characters or longer) to 2–3 characters. Strings of 4 or more characters are often compound words and phrases combining other words to represent more complex concepts. Thus, these long strings appear in the Chinese language with low frequency. In brief, site moderators could simply respond by adding all homophones of long censored keywords to a keyword ban list with little to no effect to regular users. At the same time, shortening the keywords might create confusion for readers due to missing information; however, I will show in Experiment 2 that native speakers can still infer the content of transformed weibos from shortened homophones. In my dataset, the maximum length of censored keywords is 5 characters. Therefore, I divided a long homophone in half and take either the prefix or the suffix of the homophone at random as the transformed keyword to replace the censored keyword.

4.2.3 Experiments

Experiment 1: Reposting to Sina Weibo.

To answer RQ1, I posted the transformed content weibos to Sina Weibo using multiple newly created accounts and measured the time it took for the weibos to get deleted or for the accounts to get banned. For comparison, I also posted originally censored (untransformed) weibos back to Sina Weibo and measured the same variables. I used the web interface of Sina Weibo instead of its API to post and retrieve weibos to minimize the chances of tripping automated defense systems (i.e., those systems may more aggressively filter programmatic posts arriving from API endpoints). I retrieved the list of weibos that were still published on the site every minute from a web browser session that was logged into a separate Sina Weibo account established for viewing purpose only (following the King et al. [61] method). Thus, the age of weibos has resolution at the minute timescale. The reason a viewing account was

needed is that unregistered visitors can only view the first page of another user's timeline. In order to retrieve all of the posts, I needed to access posts in other pages of the timeline. Research has shown that the majority of censored posts on Sina Weibo get censored within 24 hours of their posting [61, 124]. Relying on this result, I monitored the posts from their posting time to 48 hours after they were posted.

Experiment 2: Amazon Mechanical Turk.

To answer RQ2, I employed the online labor market Amazon Mechanical Turk (AMT) to hire native Chinese speakers to investigate if they could understand the homophone-transformed weibos. I showed the workers the transformed weibos, and provided them with the following instructions: "Please read the following post from a Chinese social media site. Some word(s) have been replaced with their homophones[2]." Participants were then asked three questions:

1. Which word(s) are the replaced word(s)?
2. Using your best guess, what are the original word(s)?
3. Did you have difficulty understanding its content?

To ensure that the workers who completed the tasks were native Chinese speakers, the instructions and questions were provided only in Chinese, accompanied with an English message asking non-Chinese speakers not to complete the task. Each HIT (Human Intelligent Task) was comprised of four weibos (asking workers to answer a total of 12 questions.) Workers were paid 20 cents for each HIT they completed, and they were allowed to complete as many HITs as they wanted, up to 152 HITs (608 weibos.) For each HIT, I obtained completed work from 3 independent workers.

[2]English translation of original Chinese instructions.

4.3 Results

Next, I report the results of two controlled experiments designed to explore RQ1 and RQ2, as well as a mathematical analysis of the likely cost a homophone scheme will impose on the current censorship apparatus (RQ3).

4.3.1 Experiment 1: Censorship effects

I created 12 new Sina Weibo accounts (excluding viewing-only accounts) for my experiment. For the purpose of reporting the results of the experiment, I define three mutually exclusive states that my accounts could fall into:

- *Active* accounts can perform all activities on the site—logging in, posting, reading other users' timeline. The viewing accounts were able to access their timelines.

- *Blocked* accounts were no longer operable. The login information of *blocked* accounts caused the site to generate the message "Sorry, your account is abnormal and cannot be logged in at this time." When my viewing accounts visited the timelines of *blocked* accounts, the message "Sorry, your current account access is suspect. You cannot access temporarily." was shown.

- *Frozen* accounts were awaiting verification. However, when cell phone numbers were provided for verification, the site always displayed the message "The system is busy, please try again," leaving the accounts in the *frozen* state and no longer operable. The login information of *frozen* accounts always lead to the verification page. Similar to *blocked* accounts, the same message was shown when my viewing accounts visited the timelines of *frozen* accounts.

Of the 12 accounts that I created, four were blocked and two were frozen, leaving six active at the end of the experiment.

For each originally censored weibo in the dataset, I posted it and its homophone-transformed version (totaling 1,216 weibos) back to Sina Weibo from accounts created for this experiment. Throughout the rest of the section, I refer to the posts I posted back to Sina Weibo as *original posts* and *transformed posts* based on their conditions. There were four progressive states that both types of my posts achieved:

- *Posted* posts are posts that were *not blocked at the time of posting*. The posters received the message "Successfully posted" from Sina Weibo when the posts were sent. *Unposted* posts caused the site to generate the message "Sorry, this content violates Weibo Community Management or related regulations and policies."

- *Published* posts are *posted* posts that my viewing accounts were able to see within 48 hours after they were posted.

- *Removed* posts are *published* posts that my viewing accounts saw at one point but disappeared from their posters' timelines at a later time within 48 hours after they were posted. However, the poster accounts were still *active*.

- *Censored* posts are *published* posts that were not visible at the 48-hour mark for any reasons, including account termination.

I calculated the age of each of the published posts from the time that I posted them to Sina Weibo to the last time the viewing accounts saw the posts. Since I defined posts to be uncensored at the 48-hour mark, I stopped checking a post after 48 hours after the time of its posting. Thus, the age of my posts was capped at 48 hours.

Keyword transformations & censorship.

Of the 1,216 weibos posted to Sina Weibo, 102 posts did not get published (8.39%): 56 original content posts (9.21%) and 46 transformed posts (7.57%). Of the posts

Table 5: Number of Weibo posts that survived through each stage of censorship.

	Original	Transformed	Total
Posts	608 (100%)	608 (100%)	1,216
Published	552 (90.79%)	576 (94.74%)	1,128
...Not Removed	521 (85.69%)	399 (65.63%)	920
...Not Censored	326 (53.62%)	337 (55.43%)	663

that did not get published, 7 original posts and 10 transformed posts were not posted (blocked at the time of posting) (4 posts from the same censored weibos.) Therefore, in total, 552 originally posts and 576 transformed posts were published, a significant difference in publishing rate ($\chi^2 = 6.219$, $p = 0.01$).

Out of the 1,128 published posts (552 original and 576 transformed,) 208 of them were removed (31 original and 177 transformed,) and 465 posts were censored (226 original and 239 transformed.) There is a significant difference in posts being removed between original and transformed posts ($\chi^2 = 116.538$, $p < 0.0001$) with transformed posts being removed more, note that transformed posts were more likely to be published than original ones. There is no statistical significance between the censorship of transformed and original content posts. Table 5 shows the number of weibo posts my viewing accounts observed after each stage of censorship. For the removed posts, the transformation of censored keywords allowed posts to last longer on Sina Weibo than the original posts ($W = 1830$, $p < 0.01$). The mean adjusted age of the removed transformed posts was 3.94 hours ($\sigma = 5.51$) and the mean for the removed original content posts was 1.3 hours ($\sigma = 1.25$), a threefold difference.

Age of weibos & censorship.

To figure out whether the original posted dates of the censored weibos also have an effect on removal of the published transformed and original posts, I accounted for the variation in the distribution of the posted dates of censored weibos in the dataset by using the ratio of between the number removed posts (transformed and original)

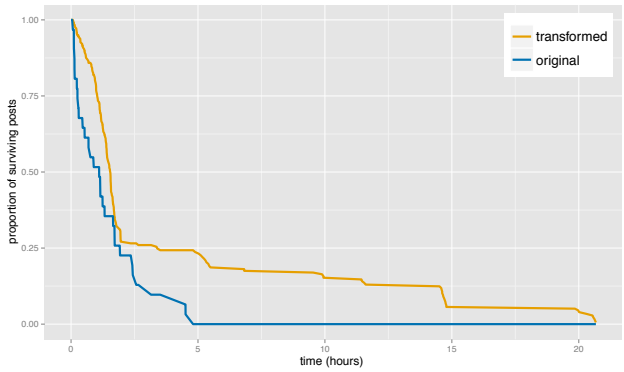


Figure 10: Proportion of *removed* posts surviving censorship, normalizing to posts' adjusted age. X-axis: Adjusted age; Y-axis: Proportion of *removed* posts.

and the number of censored weibos, based on the month the censored weibos were originally posted.

There is a significant positive correlation between the posted dates of censored weibos and the percentage of original posts removed ($\rho = 0.6478, p < 0.0001$). The correlation between the posted date and the percentage of transformed posts removed is also statistically significant ($\rho = 0.6434, p < 0.0001$).

The results of Experiment 1 show that posts with censored keywords replaced with their homophones have a higher tendency to pass through automatic keyword detection and consequently, getting published to other users and the public on Sina Weibo. While there is no significant association between posts ultimately getting censored and whether they were transformed, the age of transformed posts were significantly higher than original posts before they were removed.

4.3.2 Experiment 2: Interpretability

In Experiment 2, 22 workers completed 456 assignments. Each assignment contains 4 different transformed weibos, resulting in 1,824 *impressions* of my 608 transformed weibos. Out of 1,824 impressions, in only 52 impressions (2.85%) Tukers indicated that they had difficulty understanding the content of the transformed weibos. There were 46 transformed weibos that created confusion for 1 worker, and 3 transformed

weibos created confusion for 2 workers. There were no weibos that created confusion for all 3 workers. Table 6 summarizes the statistics of weibos and worker impressions that reported confusion.

Upon close inspection of the 3 weibos that caused 2 workers difficulties with content comprehension, 1 weibo was a reply to other weibos and had omitted some parts of the thread such as original text and images. The other 2 weibos were all originally posted in 2013, nearly 2 years prior to my study. Although these weibos were discussing current events at the time, all had important keywords of each story replaced by their homophones.

To evaluate whether the workers were able to identify the transformed keywords and the original censored keywords, I considered an answer from the workers to be correct if either (1) it was the same as the keyword, (2) it was a substring of the keyword, or (3) the keyword was its substring. Then, I calculated the portion of correct keywords as a *correctness score*. Out of 1,824 impressions, there were 617 (33.83%) that were able to detect all the transformed keywords in the weibo, and 1,200 (65.79%) detected at least half of the transformed keywords. 539 impressions (29.55%) were able to guess all the original censored keywords, and 1,091 (59.81%) were able to guess at least half of the original keywords. There were 517 impressions (28.34%) that were able to detect all transformed keywords and guessed the original words correctly. Surprisingly, 3 of them, with 3 different censored weibos, reported that they were still confused with the content of the weibos.

Logistic regressions predicting whether the workers were confused with the content of the weibos from the correctness score of both transformed keywords and original keywords show significant effects ($p = 0.03$ for transformed keywords and $p < 0.001$ for original keywords), with the correctness score for the original keywords having a steeper slope. However, the number of censored keywords and the combined length of all censored keywords do not have significant effects on the correctness scores of

Table 6: Number of impressions, weibos and workers' understanding of weibo content.

	Impressions	Weibos
Total	1,824 (100%)	608 (100%)
Confusing	52 (2.85%)	–
. . . to 1 worker	–	46 (7.57%)
. . . to 2 workers	–	3 (0.49%)
No Confusion	1,772 (97.15%)	559 (91.94%)

both transformed and original keywords, neither do they have significant effects on workers' understanding of the content of weibos.

In summary, I found that in 65% of the impressions, Turkers were able to detect at least half of the homophones of the censored keywords, and more than half of the impressions were able to guess original censored keywords themselves. The ability to identify the homophones and guess the original keywords demonstrates understanding of the content of the weibos. For 605 out of 608 of the transformed posts in my dataset, the majority of workers were able to understand the content from reading only the transformed posts.

4.3.3 Analysis: Cost to adversaries

Finally, I explored what steps the current censorship machinery (an adversarial relationship in this context, and hereafter referred to as "adversaries") would need to adapt to the technique introduced in this chapter, as well as what costs might be associated with those adaptations. As the homophones scheme introduces considerable "noise" and false positives into the weibo stream, it is likely cost adversaries valuable time and human resources. Adversaries seem likely to resort to two possible counter-measures, one machine-based and the other human-oriented. First, censors could simply add all possible homophones for a given censored term to the keyword ban list. Alternatively, censors might counter homophones with more human labor

to sort homophones standing in for censored keywords from coincidences (uses of my homophones that are not associated with censored terms). In either case, adversaries will have to deal with a potentially large number of false positives generated by my approach. Next, I analyzed how many false positive they can expect to deal with on average. In the machine-based solutions, these would amount to inadvertently blocked weibos; in the human labor case, these false positives would amount extra human labor that would need to be expended.

From the dataset of 4,441 censored weibos, there were a total of 422 censored keywords, and my algorithm generated 8,400 unique homophones that have the frequency score in the top $k = 20$. I calculated the document frequency (one weibo treated as one document) of the homophones in my public timeline corpus as a measure of how commonly these homophone phrases appear in Chinese social media. (This calculation is used as an alternative to querying the search Sina Weibo API, due to the API call limit.) My calculation may be considered the lower bound on how common the phrases are actually used in social media communication.

For each censored keyword W with the top-20 homophones $\widetilde{W}_1 ... \widetilde{W}_k$, I calculated the false positives generated by calculating the average document frequency of all homophones. In the case that W is composed of 4 or more characters, I considered the document frequency of all possible shortened keywords to be the number of false positive generated.

Then, for each censored keyword W, I calculated the average false positives generated over all of its homophones. I then calculated the average false positive generated in my dataset over all censored keywords. Algorithm 1 summarizes this process in pseudocode, the method used to calculate the number of false positive weibos for each censored keyword.

On average, each of the censored keywords matches 5,510 weibos in the uncensored corpus. The uncensored sample corpus is only a fraction of the actual posts on Sina

Algorithm 1: Estimating false positive weibos

AverageFP

Data: U: Uncensored weibo corpus
Input: W: Censored keyword
Output: $\bar{k}$: Average number of false positives for W
$k \leftarrow EstimateFP(W)$
$\bar{k} \leftarrow k/|GenHphone(W)|$
return $\bar{k}$

EstimateFP

Data: $U \leftarrow$ Uncensored weibo corpus
Input: $W \leftarrow$ Censored keyword
Output: $k \leftarrow$ Number of weibos matching W's homophones
for $\widetilde{W_i}$ *in* $GenHphone(W)$ **do**
    $n \leftarrow len(W)$
    if $n < 4$ **then**
        $S_i \leftarrow \{u \in U : u \text{ contains } \widetilde{W_i}\}$
    else
        $\widetilde{W_i'} \leftarrow \{\text{all shortened versions of } \widetilde{W_i}\}$
        $S_i \leftarrow \{u \in U : u \text{ contains any of } \widetilde{W_i'}\}$
$k \leftarrow |\bigcup S_i|$
return k

Weibo; there are approximately 100 million weibos made daily on Sina Weibo [124]. Scaling the figure above to the actual amount of weibos sent daily, the transformation would match an average of 47,000 false-positive weibos *per day, per censored keywords*. With 422 censored keywords (perhaps an under-approximation of the actual number of censored terms at work at any given time), there would be nearly 20 million false positive weibos each day, or approximately 20% of weibos sent daily.

The other option, given the current state of censorship on Sina Weibo, would be human review. Given that an efficient censorship worker can read approximately 50 weibos per minute [124], it would take more than 15 new human-hours each day to filter the false-positive weibos generated from each homophone-transformed keywords.

4.4 Discussion

First, I found that while homophone-transformed weibos ultimately get censored at the same rate as unaltered ones, they last on the site an average of three times longer than unaltered posts. It seems likely that this extra time would permit messages to spread to more people—possibly providing more time to forward the message to others. In Experiment 2, I found that Turkers who natively speak Chinese can interpret the altered message. The datasets and methods used in this chapter somewhat divorce weibos from their natural context: the weibos used here come from the past and Turkers are not the intended recipients (i.e., they don't follow the person who wrote them). Therefore, the set-up of Experiment 2 presents a relatively challenging environment for re-interpretation, one that I would argue suggests that in natural settings this method would prove highly usable. Finally, given the very large number of false positives this mechanism would introduce to the current censorship apparatus, it seems unfeasible that simply adding all possible homophones to a ban list would sufficiently address the new technique. It would interfere with too much otherwise innocuous conversation happening on Sina Weibo. (After all, Sina Weibo exists in the first place to permit this conversation to happen in a controlled space.) Rather, it seems likely that additional human effort would have to be used to counteract the technique presented here; the costs associated with that intervention appear steep, as discussed in the section above.

Turning to the results of Experiment 1, it may seem counter-intuitive that a large number of originally censored posts can now be successfully posted to Sina Weibo. There are two main explanations for this. First, the accounts that I used to post these weibos were newly created accounts without any followers. In contrast, the accounts that originally posted censored weibos were popular accounts with thousands of followers. Therefore, the adversaries might have been more lenient with my

accounts since the reach of the posts were considerably lower than those censored weibos. Second, the censored weibos were not presently topical. Some of the censored weibos in my dataset discussed events that ended long before the time I posted them back to Sina Weibo. Consequently, the posts about these events might no longer be under adversaries' watch, as we can see from the positive correlation between the original posted dates of censored weibos and the percentage of posts removed. For this reason, I measured the *relative* decrease in censorship after applying homophone transformations to my corpus.

Using homophones to transform censored keywords proved easy to understand by native speakers from the results of Experiment 2. None of the workers were confused with the content of 559 out of 608 (91.94%) transformed weibos, and the majority of the workers understood nearly all of my posts (605 out of 608 posts, 99.51%). Of course, workers need to have some background knowledge of the topics of the posts. Workers that could not identify the transformed keywords did not have an awareness of the topic nor the surrounding context. My results show a significant correlation between inability to identify transformed keywords and original keywords, and confusion with the content. It is clear that transforming censored keywords into homophones does not prohibit native speakers from understanding the content of the posts.

4.5 Limitations

For practical and ethical reasons, I did not re-appropriate existing accounts for use in my experiments. They might be compromised and potentially even endanger the people operating them. Although the Real Name Policy is not implemented on Sina Weibo [117], existing accounts might contain personal information that can be linked to real identities of account holders. Therefore, I used all newly created accounts with anonymous email addresses and, when requested for verification, anonymous cell

phone numbers to protect the safety and privacy of all parties involved. Consequently, the effects that I see in my experiments may differ in the context of well-established accounts.

4.6 Design implications & future work

The results suggest that it may be possible to construct a tool to automatically generate homophones of known censored keywords to circumvent censorship on Sina Weibo. With further engineering, all computational components in this chapter—censored and uncensored weibos crawlers, the censored keywords extraction algorithm, as well as the homophone generation algorithm—can likely be put to work together to create a tool to combat censorship in Chinese social media in real-time. Miniaturizing and scaling these technological components (for example, to live in the browser), will take effort, but is likely possible. In the next chapter, I detail the development of the real-time system, *CENSE*, that utilizes the algorithms presented in this chapter to help Chinese social media users circumvent censorship using the homophone replacement technique.

CHAPTER V

REAL-TIME SYSTEM TO CIRCUMVENT CENSORSHIP ON CHINESE SOCIAL MEDIA

Previous research and my previous work, discussed in the previous two chapters, presented promising directions towards developing viable methods of circumventing censorship on Chinese social media. However, to date, there have been no attempts to develop such systems to aid Chinese social media users in doing so.

Based on the results of my study presented in Chapter 3, censorship causes reduced participation on Chinese social media, even though the majority of Chinese Internet users do not broadcast sensitive speech that might be subject to censorship. A synthesis of the results of my interview study with Chinese social media users and of the quantitative analysis of Chinese social media data presented in Chapter 3 motivated the implementation of an automated system, *CENSE*, that will not only allow Chinese social media users to circumvent censorship but also make the censorship model transparent to users.

Figure 11 shows the interface of the front-end client of CENSE. Users of CENSE are instantly informed when their posts contain one or more censored keywords. Then, users are presented with homophone suggestions as possible replacements for those words. Furthermore, the front-end client of CENSE allows users to manually edit the suggested post to match their own preferences. Behind the front-end client, CENSE is powered by three modules in the back-end server, all of which run on a single server instance (currently hosted on the Georgia Tech network). The three modules work together to gather censored and uncensored posts from Sina Weibo, extract censored keywords from both types of posts, and generate homophones of

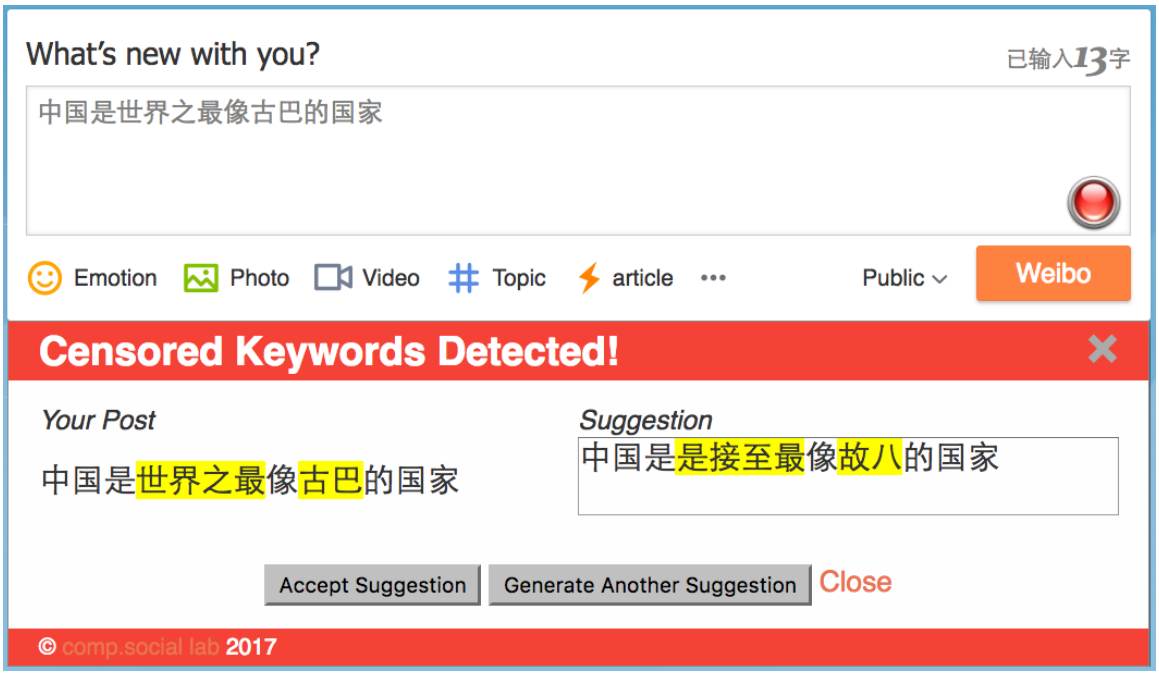


Figure 11: CENSE system integrated into Sina Weibo webpage. The system detected that the post "中国是世界之最像古巴的国家" ("China is the world's most similar country to Cuba") contains two censored keywords/phrases: *world's most (shì jiè zhī zuì)* and *Cuba (gǔ bā)*. The system then gives a suggestion to change the post to "中国是是接至最像故八的国家" ("China is connected to the most similar former eight countries") where the two censored keywords were replaced by their homophones: *connected to the most (shì jiē zhì zuì)* and *former eight (gǔ bā)*, respectively.

these censored keywords based on the algorithm presented in Chapter 4. Together, the front-end client and the back-end server function as a real-time system to aid Chinese social media users in circumventing censorship.

Before describing the implementation of the system later in this chapter, I first present the formative interview study designed to assess the opinions of Chinese social media users regarding the use of a censorship circumvention tool on Chinese social media. The details of the design and implementation of CENSE system components are then presented along with the results of user evaluation of the front-end client design. Finally, I present screenshots of a use case scenario of the CENSE system before concluding this chapter with discussions on system limitations and future work stemmed from this system.

5.1 Chinese Social Media Users' Opinions on a Censorship Circumvention Tool

As a part of the interview study regarding the usage of Chinese social media under censorship presented in Chapter 3, participants of that interview study were also asked questions regarding practices and additional tools they employ to circumvent censorship on Chinese social media. Specifically, I explored two primary research questions with this part of the study:

RQ1. What are the current techniques used to circumvent censorship?

RQ2. How would the circumvention tool, if available, be incorporated into daily social media use?

After first summarizing the research methods used in the interview study. I present the semi-structured interview questions and the results of the interviews. I then conclude this section with takeaways obtained from the interviews and describe how the interview study shaped my development of CENSE, a censorship circumvention tool.

5.1.1 Methods

Participants in the interview study were the same group that participated in the other interview study presented in Chapter 3. To summarize, participants were carefully selected based on the following criteria to minimize risks associated with participating in the study due to the sensitive nature of the interview topic:

1. Participants were Chinese citizens to ensure that they were familiar with the culture, politics, and Internet of China.

2. Participants were located in the US to facilitate in-person or domestic-US phone call interviews to mitigate risks and protect participants' identities and security.

3. To ensure that they were familiar with Chinese social media, participants were users of Chinese social media.

During the interviews, participants were first asked questions designed to answer the research questions presented in Chapter 3. Next, they were asked questions aimed at answering the research questions presented in this section, i.e., ones that involved their knowledge of censorship circumvention techniques and their opinions on automated censorship circumvention tools. The interview questions based on the research questions presented in this section are presented next.

Additional details of the participant recruitment process and the interview sessions are described in Chapter 3.

5.1.1.1 Interview Questions

1. Are there any techniques that you use to circumvent censorship? What are they?
   (a) Where did you learn those techniques?
   (b) Do you observe other people using the same techniques?
2. How effective are the techniques that you use in your experience?
3. How do you adapt the techniques to make sure that you keep up with new censorship practices?
4. If there were a tool that automatically warns you that your posts could be censored before you post them, would you use it? How would you use it?
   (a) How effective do you think the tool would be in helping you circumvent censorship?
   (b) Would you participate more on the sites?
5. If the tool also suggests that you change some words to avoid censorship, would you use this tool? How would you use it?

(a) How effective do you think the tool would be in helping you circumvent censorship?

(b) Would you participate more on the sites?

5.1.2 Results

The results of the interview study are grouped into two sections. First, censorship circumvention techniques reported by the participants are presented to provide the context of participants' awareness of censorship circumvention techniques. Then, participants' opinions on an automated censorship circumvention tool are reported.

5.1.2.1 Censorship Circumvention Technique

Nearly all of the 11 participants admitted to knowing techniques designed to circumvent censorship on Chinese social media. Specifically, the two techniques mentioned by the participants were word substitution (e.g. replacing sensitive words with homophones) and word manipulation (e.g. inserting symbols between characters in a word).

> *Using generic name or a different name. Chinese characters have a lot of things sound alike. For example, a lot of people on Sina Weibo are now calling Jiang Zhemin "toad." If you see that then it's definitely mentioning him by the name. (P1)*

> *In Chinese social media, if you write a word, you can write comma in between the words and that doesn't get censored, and people can still understand. (P2)*

Despite their awareness of censorship circumvention techniques, almost all participants had never used these techniques themselves for two main reasons. First, the participants did not post content involving sensitive topics on Chinese social media. Thus, they did not see the need to replace sensitive words when posting. Second,

the participants usually did not know what words were considered "sensitive" at any given time. Therefore, they could not identify which words in their posts to replace.

5.1.2.2 Usefulness and Effectiveness of an Automated Censorship Circumvention Tool

When the concept of an automated tool to circumvent censorship was presented to them during interviews, participants were asked to evaluate the usefulness of the proposed tool. Five out of 11 participants stated that they believed the tool would be useful if it were not too intrusive.

> *Maybe it will be helpful before you add little characters between the words, you know which words will be censored, you know where to add the characters. (P2)*
>
> *I guess it would be like a spell check so I would see what I'm writing about is OK if it is then I would post it. (P5)*

These five participants who stated they would find the tool useful predicated its effectiveness in circumventing censorship on its ability to detect sensitive keywords, and on whether post contents were understandable despite the noise created by the tool's manipulating the post's words.

> *If the tool knows all the censorship of sensitive words then it will be effective. I hope that the government won't notice this tool otherwise they might block this tool also. (P7)*
>
> *I think that would be more effective than using the same derived words but the potential problem is if the readers will be able to understand because it's kind of an encrypted message. (P5)*

Four out of 11 participants explicitly stated that they did not think an automated tool to help circumvent censorship on Chinese social media would be helpful. These

participants raised three issues in support of this view. First, two participants raised the concern that the tool would emphasize the limitations of freedom on Chinese social media. Having a tool that explicitly warns social media users that their posts might be censored could make users angry and drive them away from social media platforms.

> *I think it will make people feel like we are not free in China... I don't like this feeling. Once you type something and it tells you that your words are sensitive on Sina Weibo, it makes me feel bad that I won't use Sina Weibo anymore. The tool will make me want to post less on Sina Weibo. (P6)*

The second issue was related to users' trust in the tool. Two of the four participants expressed concern that the tool would not be useful because they could not trust it. People might suspect that the tool was released by the Chinese government as simply another way to suppress speech.

> *I think this kind of tool will make people angry. They will think the tool is made by the government. People don't want to be limited. (P6)*
>
> *I don't think that will be useful because people know they will get censored. I don't trust those tools because it will tell me I will get censored but how do those tools know I'm getting censored. They can't get into the mind of the manual censors. (P11)*

Finally, one participant said that he/she did not find the tool useful because he/she would never post something of a sensitive nature on Chinese social media.

> *I don't think it will be useful to me personally because I don't post these things. I feel posts getting deleted is far away from my life, so I wouldn't feel like it's too useful for me. I would be curious how it works and see but using it in a daily basis, I don't think so. (P10)*

The other two out of 11 participants refrained from commenting on the usefulness of the tool until they could see it in action.

The next subsection explains how the results of the interview study helped inform design decisions of CENSE, the censorship circumvention system that I developed.

5.1.3 Design Implications

The results of the interview study showed that most participants who are Chinese social media users believe that a censorship circumvention tool would be helpful in their social media routine if the tool does not interfere with their social media experiences.

Comments from the participants who did not find the tool useful were used to inform design decisions to ensure users' trust in the tool and to ensure that the tool encourages rather than suppresses user participation in Chinese social media.

Thus, the interview study informed a few key design elements of the censorship circumvention system:

- The system should blend seamlessly with Chinese social media.
- The system must be unobtrusive.
- The system should encourage user participation in Chinese social media to gain users' trust.

In the next sections, I detail the technology behind the CENSE system and how I incorporated the design takeaways listed above into the development of the system.

5.2 System Components

The system CENSE consists of two main components: a back-end server and a front-end client. The back-end server utilizes the homophone transformation algorithm

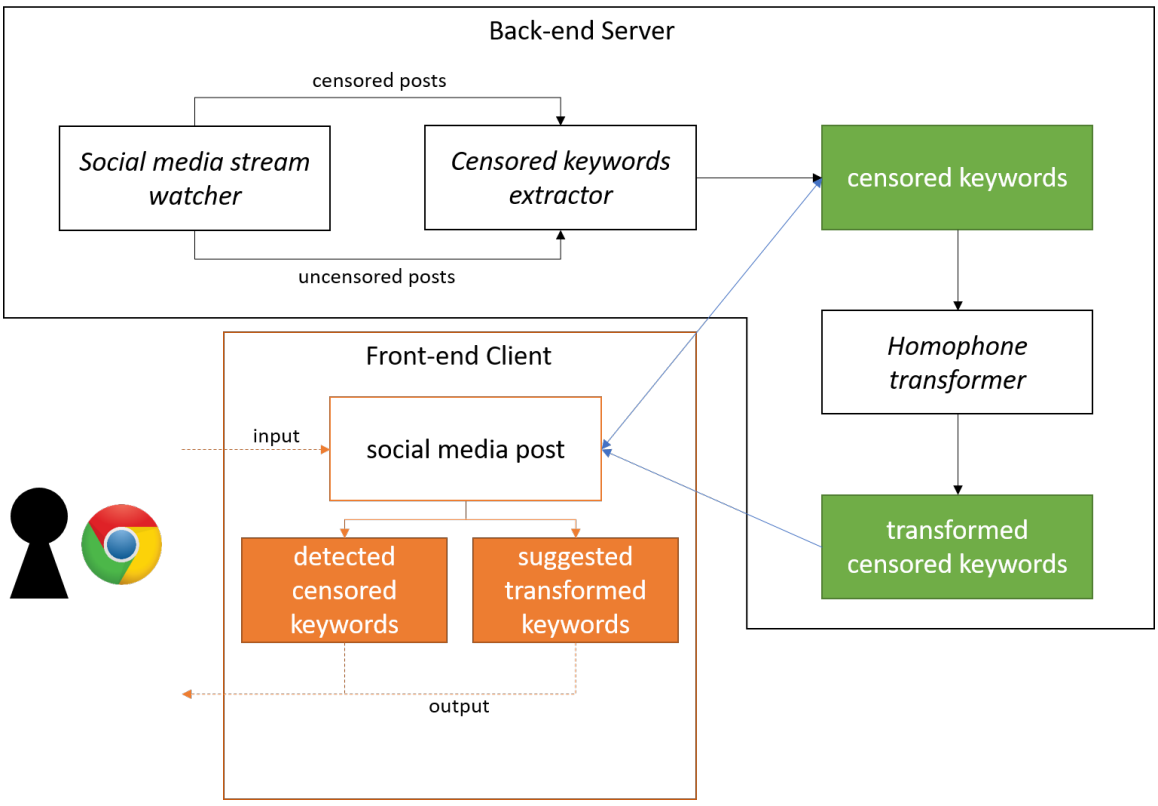


Figure 12: Components of the CENSE system.

presented in Chapter 4 that I created to detect censored keywords and suggest corresponding homophone replacements. The front-end client interacts with users, allowing them to replace detected censored keywords on their social media posts with homophones.

Figure 12 shows interactions among the different modules in the back-end server, interactions between the front-end client and the back-end server, and the interactions between users and the front-end client.

The next two sections detail the design of the back-end server and the front-end client components.

5.3 Back-end Server Design

The back-end server component of CENSE facilitates the system's logical operations, i.e. the back-end server handles the data mining process that extracts current censored keywords on Chinese social media. Additionally, the back-end server finds homophone replacements for censored keywords. This back-end server consists of three modules: the Social Media Stream Watcher, the Censored Keywords Extractor, and the Homophone Transformer. Figure 13 shows how the data flow between these different components of the back-end server.

5.3.1 Social Media Stream Watcher

In order for the system to be up to date on current keywords that are prone to censorship, it must constantly monitor both censored and uncensored posts on Chinese social media, and use this data to detect censored keywords. In my previous work described in Chapter 4, both censored and uncensored posts were only obtained once from a Chinese social media site and a Chinese social media curator. Extending this previous work, the new module was developed to constantly monitor a Chinese social media stream of uncensored posts (in this case, the public timeline of Sina Weibo) and a stream of known censored posts (in this case, censored posts on Sina Weibo curated by Weiboscope).

Since the posts from Sina Weibo's public timeline stream are already published on the platform, they have already passed the automated filter. My system compares the posts from Sina Weibo's public timeline to censored posts in order to identify censored keywords. To monitor for Sina Weibo uncensored posts, I used the Sina Weibo Open API to monitor the public timeline stream. Sina Weibo Open API has a rate limit of 150 requests per hour [68], and each request to query the public timeline returns upto 200 posts. Thus, in one day, I could obtain at most 720,000 uncensored posts from Sina Weibo (0.72% of 100 million posts made daily). The posts from Sina

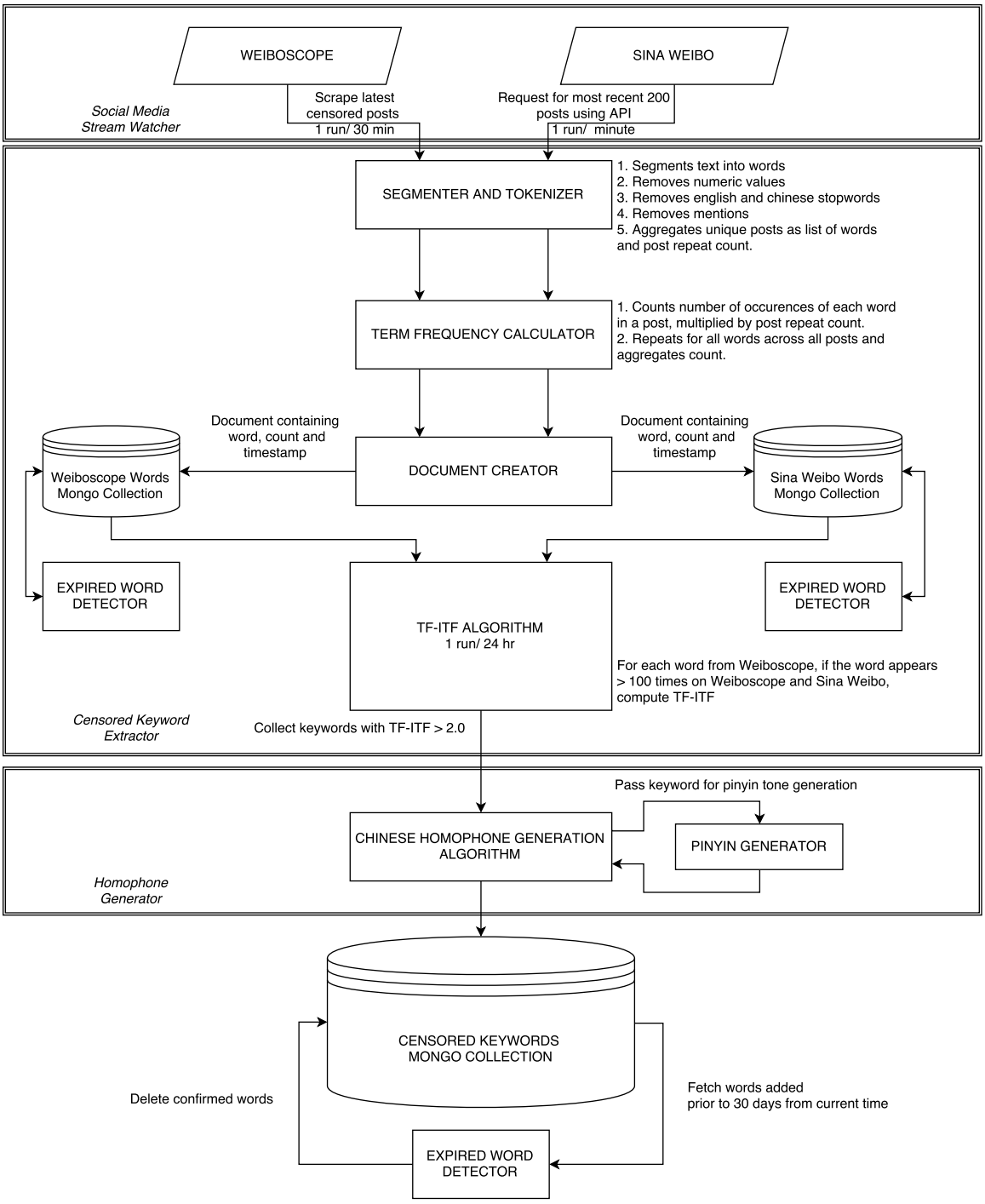


Figure 13: Diagram showing the flow of data between components of the back-end server.

Weibo's public timeline are kept for 30 days to serve as the *uncensored* corpus for use by the Censored Keyword Extractor module.

Identifying censored posts on Sina Weibo has proven to be a challenging task due to Sina Weibo Open API's rate limit. Fortunately, Weiboscope [38] extensively monitors popular accounts on Sina Weibo for censored posts and makes this data available to the public. My system continuously obtains Sina Weibo censored posts from Weiboscope to compare with uncensored posts obtained from Sina Weibo's public timeline. The number of daily censored posts on Sina Weibo is minuscule compared to the number of uncensored posts obtainable from Sina Weibo's public timeline. Weiboscope detects approximately 40-50 censored posts per day. However, keywords that are prone to be censored could remain in the filtering system for an extended period of time. Therefore, censored posts from a 180-day period form the *censored* corpus for the Censored Keyword Extractor module.

5.3.2 Censored Keyword Extractor

Like the censored keywords extraction algorithm presented in Chapter 4, the Censored Keyword Extractor module utilizes the term frequency-inverse document frequency (*TF-IDF*) technique to extract keywords that appear frequently in the censored corpus but rarely in the uncensored corpus.

Two, three, and four-character words/phrases are extracted from the censored corpus. Here, I decided to exclude words/phrases that are more than four characters long because they are not common in the Chinese language. Then, terms that appear less than 100 times (note that this number is significantly higher than that used in the technique discussed in Chapter 4 due to the larger datasets employed here) in the combined censored and uncensored corpora are removed to ensure that those terms that remain are ones that commonly appear on Chinese social media. With each post in the uncensored corpus considered to be a document, the TF-IDF score for each

term from the censored corpus is then calculated. Thus, for a term w:

$$\text{TF-IDF}(w) = \frac{\text{frequency of } w \text{ in censored corpus}}{\#\text{ of posts in uncensored corpus containing } w}$$

Then, terms with TF-IDF scores greater than 2.0 are assumed to be censored keywords. Since the stream of data that feeds into the Censored Keyword Extractor is now dynamic, I decided to not include any assumed ground truth about words or phrases that could be censored keywords as I did with Freeweibo's Hot Search keywords in Chapter 4.

5.3.3 Homophone Transformer

Using the same mechanism as in the homophone generation algorithm discussed in Chapter 4, the Homophone Transformer module takes in censored keywords generated from the previous module. Then, for each keyword W, its homophones $\widetilde{W_1}, \widetilde{W_2}, \ldots, \widetilde{W_n}$ are generated from characters having the same root sounds (ignoring tones) but in different forms. Then, each homophone $\widetilde{W_i}$ is assigned a score that is the sum of the frequency percentile of the characters. To ensure that the homophones generated by the system are ones that commonly appear in the Chinese language, a homophone with a score in the 20 highest scores is randomly selected to be suggested to the user as a replacement for the original keywords.

5.4 Back-end Server Implementation

All three modules comprising the back-end server—the Social Media Stream Watcher, the Censored Keyword Extractor, and the Homophone Transformer—were implemented using Python 2.7. All data are stored in MongoDB, a NoSQL database service. Currently, the back-end server is deployed on a server instance on the Georgia Tech network. The back-end server is currently running on a machine with a 2.4GHz 16-core processor, 96GB of RAM, and gigabit link to the Internet.

5.4.1 Social Media Stream Watcher

The Social Media Stream Watcher obtains the latest posts, up to 200 a minute, from Sina Weibo's public timeline through the Sina Weibo Open API. After these posts are segmented into Chinese words using Jieba Chinese text segmentation [102], individual words are then incorporated into the *uncensored* corpus and stored in the database.

Additionally, the Social Media Stream Watcher obtains the latest censored posts from Weiboscope, the Chinese social media curator site. However, since there are not as many censored posts on Weiboscope as there are posts on Sina Weibo public timeline, the Social Media Stream Watcher only obtains data from Weiboscope every 30 minutes. After the censored posts are obtained, all two-, three-, and four-character words are extracted and stored in the *censored* corpus on the database. I did not employ Chinese word segmentation on the censored posts since lexicon variations such as memes, morphs, and homophones that could already be in use, and word segmentation might misinterpret these variations as parts of other words.

For both the censored and uncensored corpora, the database stores words in the appropriate table and, along with the words themselves, their total and daily occurrence counts. Words in the uncensored corpus (i.e., from the Sina Weibo's public timeline) expire 30 days after the day they were obtained, while words in the censored corpus (from Weiboscope) expire 180 days after the day they were obtained. Once words expire, they no longer contribute to the total occurrence count. I decided to expire words in the censored corpus much later than those in the uncensored corpus because otherwise there is not enough data in the uncensored corpus to detect censored keywords.

5.4.2 Censored Keyword Extractor

The Censored Keyword Extractor compares unexpired words from the censored and uncensored corpora in the database to detect censored keywords on Sina Weibo.

Using the total count of words in the database, the Censored Keyword Extractor can calculate the TF-IDF score for each word in the censored corpus. To ensure that the censored keywords are ones that appear commonly on Chinese social media, only the words in the censored corpus that appear more than 100 times (total count > 100) in the combined censored and uncensored corpora are assigned the TF-IDF score. Then, I apply the threshold of TF-IDF(w) ≥ 2.0 to ensure that the censored keywords appear frequently enough among the censored posts. Finally, the keywords are stored in the database as censored keywords.

The Censored Keyword Extractor executes an extraction once per day to obtain censored keywords that have appeared on Sina Weibo within the past 24 hours. While the extractor module can be executed more frequently to catch censored keywords as they appear on Sina Weibo, my experiments have shown that a 24-hour period is currently the optimum time period over which to gather enough censored posts from Weiboscope to run through the TF-IDF algorithm because the number of censored posts available through Weiboscope is limited.

Once extracted, homophones of censored keywords are generated through the Homophone Transformer module, which generates the top 20 homophones that are most likely to appear in the Chinese language. Censored keywords and their homophones are then maintained in the database of the CENSE system for 30 days.

5.4.3 Homophone Transformer

The Homophone Transformer generates homophones of censored keywords extracted from the Censored Keyword Extractor module. Currently, the Homophone Transformer module utilizes Da's character frequency list of Classical and Modern Chinese [25] to calculate the probability of a Chinese character appearing in a Chinese

language text. The homophones are generated using the algorithm presented in Chapter 4. Once all homophones are generated, the homophones with the 20 highest frequency scores are stored in the database of the CENSE system. When the front-end client requests a homophone, one of the stored homophones is randomly selected as a replacement suggestion.

In future work, Da's character frequency list of Classical and Modern Chinese can be replaced by the frequency list of Chinese social media gathered from the Social Media Stream Watcher module.

5.4.4 API Endpoints

To facilitate communication between the back-end server and the front-end client, I also developed API endpoints on the back-end server side. Two endpoints are needed in order for the front-end client to function properly.

- */keywords* provides the current list of censored keywords stored in the database on the back-end server. The front-end client utilizes the *keywords* endpoint to retrieve this list, which the front-end client uses to determine whether the user-supplied post contains any censored keywords.

- */homophone* takes a word or a list of words as its input and provides the homophone suggestions of these words as outputs. Once the front-end client detects one more more censored keywords, it queries the */homophone* endpoint for homophone suggestions to replace these words. The */homophone* endpoint return one of the top 20 homophones with highest frequency scores for each keyword as possible replacements of the keywords.

I developed the API using the Python programming language and Flask web framework, which had the advantage of also being written in Python, the same programming language that I used to develop all modules of the back-end server. Consequently, I was able to quickly develop the API endpoints by repurposing the code

of the back-end server.

The API endpoints are currently served through the uWSGI engine, an interface between web servers and web frameworks for the Python programming language, on a server hosted on the Georgia Tech network. As the API endpoints are lightweight and require little processing power and bandwidth, they can be deployed in the same instance as the back-end server.

5.4.5 Performance

The Social Media Stream Watcher operates at an efficient performance. Each run of the stream watcher involves scraping the Sina Weibo's public timeline or Weiboscope, segmenting posts into words, and storing the resulting words in the database, a process that takes approximately 9 seconds to complete. The Weiboscope stream watcher is slightly faster due to the shorter latency between Georgia Tech's server and Weiboscope's server.

Together, the Censored Keyword Extractor and the Homophone Transformer take approximately one hour to execute. Since these two modules are executed only once per day, this execution time does not have a significant impact on the performance of the whole system.

The API endpoints served through the Flask framework and the uWSGI engine perform relatively well and without apparent latency. In my experiments, all calls made to the API endpoints were returned within seconds. Calls to the API endpoints did not cause any lags in the front-end client.

5.5 Front-end Client Design

To facilitate the use of the back-end server component, a front-end client was developed. The front-end client performs two major tasks. First, it automatically monitors the content of the user's social media posts. Second, it informs the user whether a post contains any censored-sensitive keywords and offers homophone suggestions as

replacements for those keywords.

5.5.1 User Interface Design

Based on the results of the formative interview study presented earlier, the design of the user interface of CENSE needs to be unobtrusive and incorporate seamlessly with users' Chinese social media experiences. Therefore, I chose to implement the front-end client as a Google Chrome browser extension, which users can easily install on their personal computers. I propose two user interface designs for the web browser extension as shown in Figure 14.

In the first interface option (Design A), a user is shown two versions of the post side-by-side: an original version and a suggested version with censored keywords are replaced by their suggested homophones. The second interface design option (Design B) displays the detected censored keywords and their homophone replacements alongside the content of the original post.

To evaluate which design of the front-end client interface best streamlines users' social media experiences, I conducted a user study to evaluate these design proposals.

5.5.2 Design Evaluation

The user study I conducted employed a think-aloud protocol with four participants to evaluate the design proposals of CENSE's front-end client. I recruited participants who were at least 18 years old and are familiar with social media, although not necessarily Chinese social media, to participate in the study. Participants were recruited through advertisements such as flyers posted on the Georgia Tech campus, mailing lists of Georgia Tech students, and classified ads. Moreover, a snowball-sampling technique was also used to recruit additional participants through referral by existing participants and personal contacts of researchers. All participants were required to come to a lab location on the Georgia Tech campus at the time of the study. After participants were asked to complete a survey regarding their social media usage, they

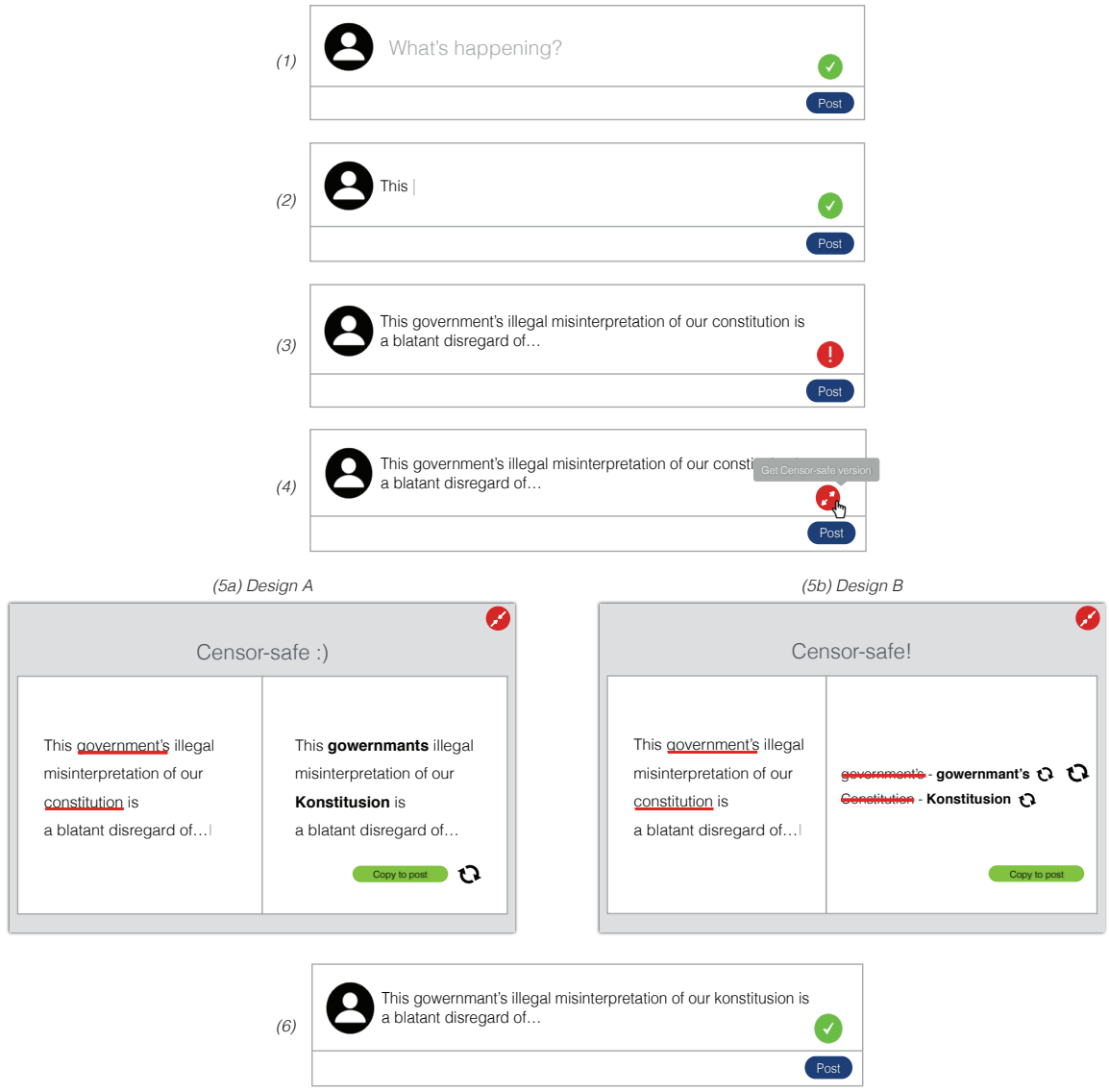


Figure 14: Two designs of the front-end interface.

were randomly assigned one of two study conditions. The participants assigned to condition A first viewed Design A and then Design B. Conversely, participants assigned to condition B first viewed Design B and then Design A. Participants were encouraged to talk out loud their opinions and thoughts in reaction to each design proposal. Participants were also asked the following questions, which were designed to evaluate the usefulness of and their preferences for each design:

1. What was the difficulty to avoid posting censored-sensitive content? Was there anything confusing or complicated?

2. How was the experience of posting on social media with this design? Why did you like (or dislike) it?

3. How was the experience of receiving the feedback from the design? Why did you like (or dislike) it?

4. Which design do you prefer? Why do you prefer your choice of design?

Because CENSE deals with the sensitive issue of Internet censorship and censorship on social media, I was also interested in determining the participants' willingness to use the system to interact with their social media profiles. In addition to questions intended to evaluate the usefulness of the design and participants' design preferences, participants were also asked questions regarding the trustworthiness of the design and of the overall system:

1. Would you use this system to circumvent censorship? Why or why not?

2. What aspects of the design or your experience with it made you trust/distrust the system?

5.5.3 Evaluation Results

In this subsection, I present the results of the user evaluation of the interface design. These results are grouped into four categories. First, participant backgrounds and censorship experiences are reported to provide context with respect to their familiarity with censorship and content removal on social media. Second, participants' evaluations of the design are presented. Then, participants' perceptions of the tool's usefulness and effectiveness are reported. Finally, I present trust in the tool as reported by the participants and their willingness to use the tool in their own social media routines.

5.5.3.1 Censorship Background

Because I had not specifically recruited only Chinese social media users to participate in this study, I employed the first few questions of the interview study to assess participants' understanding of the practice of content removal on social media. All participants expressed an awareness of censorship practices, and some participants who were from China and its governing territories had heard of or experienced censorship on Chinese social media.

> *There is stuff like this in China. It's very serious. Like Sina Weibo has important keyword flags, so it's not allowed. It's not really about posting against the government, it's basically the use of sensitive words. (P1)*
>
> *I follow activists/political figures and I know they experience it. (P2)*

5.5.3.2 Design

In all sessions, the random assignment of the order of presentation of the two user interface designs had no impact on the reactions and opinions that the participants expressed on the two designs.

Overall, participants preferred Design A over Design B because it presented the entire post content after censored keywords were replaced with their corresponding homophones. Design A allowed the participants to see how the final post would look and allowed participants to easily compare the original and the new, suggested post.

> *[Design A] allows preview so gives me the option to see if there's any hilarious effects. (P1)*
>
> *I prefer [Design A]; [Design B] is more explicit. [Design A] was more visually balanced. I'm used to having 2 things (left and right), [it looks] more balanced. (P4)*

Nevertheless, participants pointed out a few flaws of design A, especially when the post is long. First, the participants commented that design A does not allow for customization of single words. For instance, P1 stated that, in some cases, he/she would want to "come up with the words [himself/herself]" to ensure that the final post does not sound awkward and that it conveyed the meaning he/she intended. Second, showing the two full-length posts required that participants read through the suggested posts in their entirety. This process was "too much work for long posts" to ensure that they sounded fine (P2).

5.5.3.3 Usability and Effectiveness

All participants found the tool useful regardless of their opinions on the interface design. Specifically, participants found two aspects associated with the tool useful. First, it could make users who are not already aware of censorship on Chinese social media become aware that certain keywords they use are being censored. Second, the tool could help automate the process of replacing words in a post, especially a long one, to circumvent censorship.

> *[The tool is] useful to someone who isn't aware and would care to fix it with tools like this. (P4)*

Very useful, of course. For some posts that have so many sensitive words. It may be cumbersome to replace each word. (P1)

Moreover, all participants found the tool to be effective in circumventing censorship on Chinese social media, although some participants stated that the tool would not completely circumvent the entire censorship process.

This might keep the post up a day more. They are super fast about adding new keywords. I'm sure this will help but just not sure how much. Definitely net positive effect but not sure how large. (P2)

One pitfall of the tool that P1 pointed out was that the words suggested as replacements for the censored keywords could also be on the censorship agents' censored keyword list.

5.5.3.4 Trust in the Tool

All participants stated that they would have reservations about using the tool if they randomly discovered it on the Internet. Participants would want to do some research concerning its developers and purpose. In a real-world setting, participants believe that the tool would not be allowed to be released in China. Thus, if the participants were to find this tool available while they were in China, they would be more careful in adopting it.

I just question if it's allowed for this plugin to exist [in China]. Especially if it's tied into social media. But if it's separate, another platform, I could try it out and see what is sensitive in a post. (P1)

If this is a random tool [I find on the Internet], I would have to do some research first. (P2)

Because the study was conducted in a research setting, once participants were able to verify that the censorship circumvention tool was developed by a research group at Georgia Tech, they all stated that they would use the tool once it becomes available.

> *If something I post is sensitive and someone tries to prevent me I will use this to check how I could. I wouldn't want anything preventing me from posting. (P1)*
>
> *I would but just because of my deep distrust of the government. I would be skeptical of how long it would live. If it gets popular, what if it gets banned? (P2)*

5.6 Front-end Client Implementation

Based on the feedback received from the participants in the design evaluation study, I chose the major design elements from Design A to implement as the CENSE's front-end client. However, I did incorporate a few design changes to Design A based on the feedback I received from participants:

1. Allow edits of the suggested post
2. Explicitly identify the developers of the tool

The front-end client was developed as a Google Chrome extension because users can easily install this extension on their web browsers. In addition, Google Chrome extensions make heavy use of Javascript which can be easily adapted for other platforms.

Figure 15 below shows screenshots of the front-end client in the use case where the user's post contains censored keywords and the system suggests an alternative post that replaces censored keywords with corresponding homophones. In this scenario, the user interaction with the client is as follows:

1. The user logs in to his/her Sina Weibo homepage. (Figure 15a)

    (a) As the Sina Weibo homepage is loading, the CENSE client prefetches the list of censored keywords from the back-end server.

2. The user begins typing his/her post into Sina Weibo's input box. (Figure 15b)

3. Within 500 milliseconds (the delay I imposed to permit the user to finish typing), if the post does not contain any censored keywords, the CENSE client injects a green light indicator into the Sina Weibo input box, telling the user that his/her post should be safe from censorship adversaries. (Figure 15b)

4. On the other hand, if the system detects one or more censored keywords in the user's post, the CENSE client injects a red light indicator into the Sina Weibo input box within 500 milliseconds of the user finishing typing. (Figure 15c)

    (a) At the same time, the client fetches homophone replacements of the censored keywords from the back-end server to populate the suggestion supplied in later steps.

5. When the user hovers the mouse cursor over the red light indicator, this indicator transformed into the expand icon, the mouse cursor transforms into the select cursor to indicates clickability, and the tooltip is shown as a hint to the user to click on the icon. (Figure 15d)

6. When the user clicks on the hover icon, a panel instantly appears below the Sina Weibo input box. This panel contains two versions of the post: the original version obtained from the Sina Weibo input box and the suggested version with censored keywords replaced by corresponding homophones. In both versions of the post, the censored keywords and their homophone replacements are highlighted. (Figure 15e)

7. The user can edit the suggested post in the expanded panel to match his/her preferences. (Figure 15f)

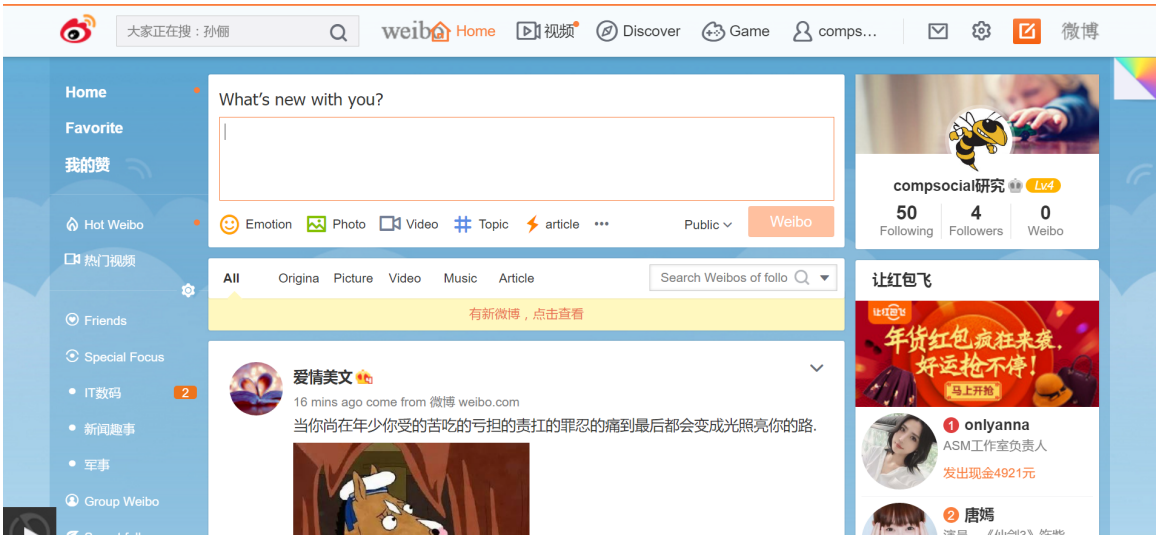
8. Alternatively, the user can ask the interface to generate a new set of homophone replacements by clicking on the "Generate Another Suggestion" button. Once this button is clicked, the suggested post will be repopulated with the new version shown in the suggested post with homophones replacing censored keywords. (Figure 15g)

    (a) The front-end client fetches the new set of homophone suggestions from the back-end server once the user clicks on the "Generate Another Suggestion" button. Although this could introduce some latency into the user experience, for the back-end server currently running on the Georgia Tech network, latency is unnoticeable.

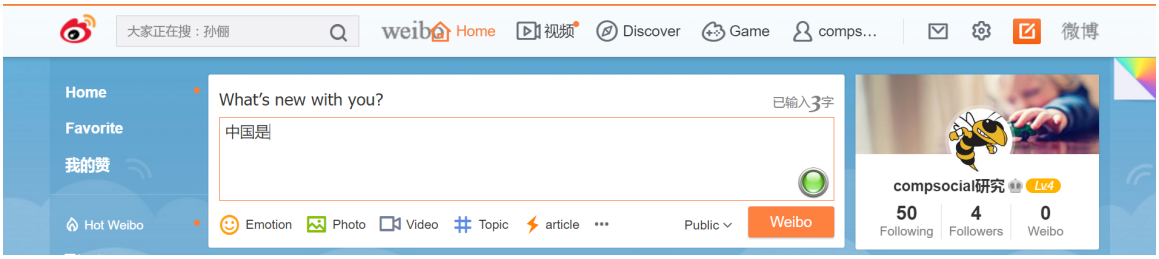
9. Once the user is finished with editing the suggested post, the user clicks the "Accept Suggestion" button. The post in the Sina Weibo input box is replaced with the suggested post (or the edited version of the suggested post). The user can then post the post to his/her timeline by clicking the "Weibo" button of Sina Weibo's interface.

5.6.1 Performance

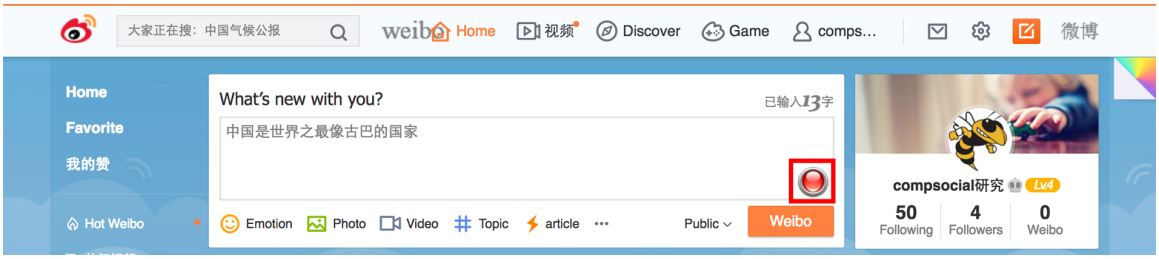
The performance of the front-end client is largely based on the performance of the back-end server and the latency between the machine that the front-end client is running on and the back-end server instance. As mentioned in the previous section, the performance of the back-end server does not introduce lags into the performance of the front-end client. Moreover, since I am running the back-end server on the Georgia Tech network and am testing the front-end client on a machine that is also part of the Georgia Tech network, there is little to no latency between the front-end



(a) Sina Weibo home screen with status update input.

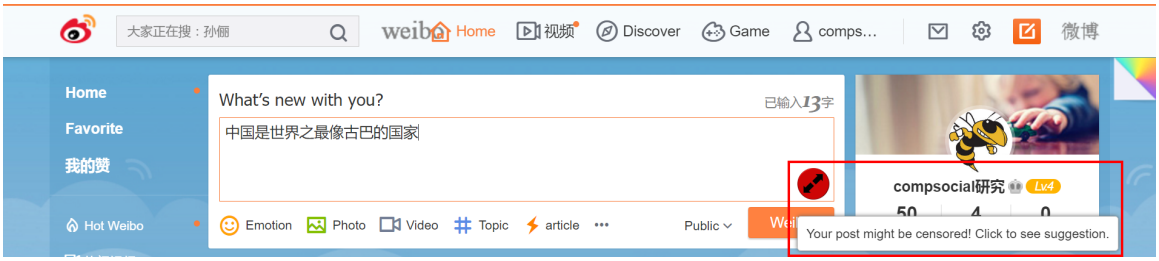


(b) As a user types in the status update box, an indicator that is part of the CENSE extension will appear at the bottom right of the textbox to indicate whether the post contains any censored keywords. A green indicator means that the post contains no censored keywords.



(c) The indicator turns red if the post contains censored keywords.

Figure 15: Screenshots of the CENSE front-end interface as a Google Chrome extension running on the default interface of Sina Weibo homepage.

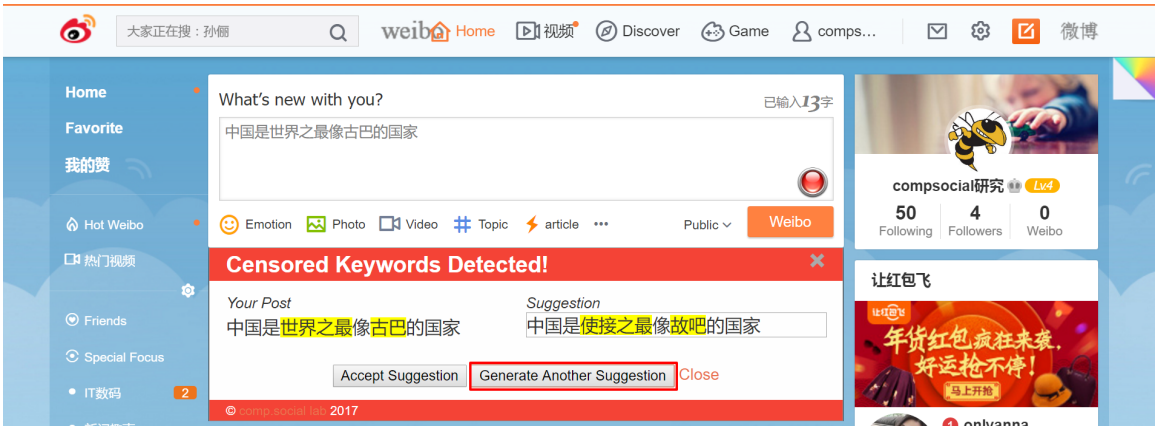


(d) Once the red indicator is hovered, the icon changes to the expand button to give the user feedback, and a tooltip is displayed to give the user the option to request additional information and instructions.

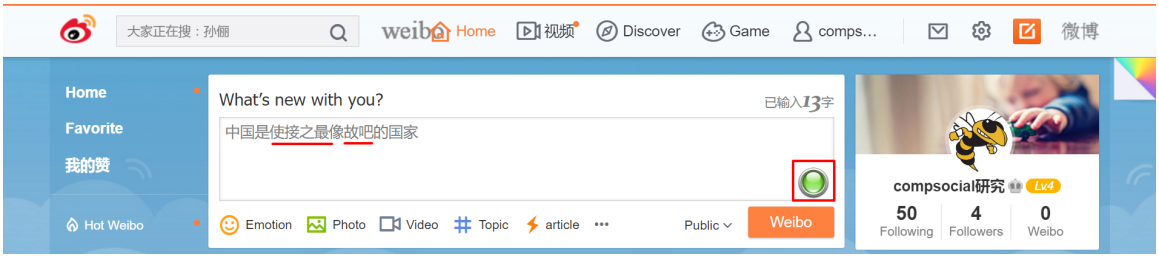
(e) When the expand button is clicked, a panel appears below the status update box to show the user's post and the suggested alternative. The censored keywords and the replaced homophones are highlighted for easy detection.

(f) Users can change the suggested post in the textbox if they do not like the suggested post.

Figure 15: (continued) Screenshots of the CENSE front-end interface as a Google Chrome extension running on the default interface of Sina Weibo homepage.



(g) Alternatively, users can request the system to generate another set of homophones to replace the censored keywords.



(h) Once the user accepts the suggestion, the new status update appears in the status update box, and the indicator turns green to show that the status no longer contains censored keywords.

Figure 15: (continued) Screenshots of the CENSE front-end interface as a Google Chrome extension running on the default interface of Sina Weibo homepage.

client and the back-end interface, resulting in near-instantaneous user interactions with the front-end client.

However, if a user is running the front-end client on a machine that has high connection latency to the back-end server, the user experience might not be as seamless. Nevertheless, since I designed the front-end client to prefetch most of the data needed for censored keyword detection and keyword replacement suggestions while the Sina Weibo homepage is loading, the user should experience little delay in interactions with the CENSE front-end client.

5.7 Evaluation

The ideal method to evaluate the effectiveness and usefulness of this system is to conduct a deployment study where the system is distributed to participants who are actual Chinese social media users. Participants then use the system for a period of time: one month, for example. Afterwards, participants take part in an interview session or a survey to evaluate two main aspects of the system, its effectiveness and usefulness. However, I did not conduct such a study as a part of this thesis. In this section, I explain the reasoning for forgoing the deployment study as well as an alternate study that can potentially be conducted to evaluate the system.

5.7.1 Arguments against a deployment study

A deployment study is a traditional method in HCI to evaluate the usefulness and effectiveness of a system. However, in the instance of my system, a deployment study is not a suitable option due to the sensitivity of the topic of censorship on Chinese social media. In a traditional user study protocol, in order for participants to participate in the study, they must first have contact with the researchers during the recruitment process, after which they download and use the system for one month. After a month, the participants contact the researchers again to complete a survey or participate in an interview session. This protocol presents many opportunities for

communications between participants and researchers to be intercepted by a third party.

In normal user study protocols where risks involved are not greater than those present in everyday life, the traditional protocol presents no potential for extraordinary risk to participants. However, in this study, the risks associated with participants intentionally circumventing censorship on multiple occasions could become a major concern, because participants in the study are most likely Chinese citizens and currently reside in China. The Chinese government has deployed advanced mechanisms not only to censor but also to monitor its citizens' Internet usage [19, 104]. Thus, using the traditional protocol for this study would expose participants to risks potentially outweighing study benefits. Therefore, I decided against performing a traditional deployment study.

Nevertheless, the CENSE system has been released as an open-source software under the MIT license for public use at `https://github.com/compsocial/CENSE`. Moreover, the usage log and user data are not being collected to preserve users' privacy and security.

5.7.2 Alternate plan for system evaluation

With the sensitivity of the topic of censorship on Chinese Internet and the risks associated with participating in a deployment study of a censorship circumvention tool in mind, I propose an alternative plan for evaluating the CENSE system. The CENSE system can be evaluated for its usefulness and effectiveness through a lab study to avoid the risks associated with posting sensitive content to Chinese social media.

The lab study would simulate the experiences that participants have to go through when posting sensitive content on Chinese social media, except that the participants will not actually posting to Chinese social media. Instead, participants interact with

a mock-up version of Sina Weibo, which replicates its functionality and interface. This way, there is no risks associated with posting sensitive content to Chinese social media in this lab study.

In this lab study, participants will be recruited to come to the lab space during the time of the study. Participants should be Chinese social media users who are familiar with Chinese social media platforms and their censorship policy. When participants arrive at the lab and agree to the consent, the participants will follow the following study protocol:

1. Participants are briefed that throughout the study, they will interact with a mock-up version of Sina Weibo, and no data generated during the study will be transmitted through the Internet to Sina Weibo.

2. Participants are encouraged to talk out loud any thoughts regarding their interactions with the mock-up Sina Weibo that occurred during the study.

3. The participants are presented with a mock-up version on Sina Weibo on a lab machine.

4. Participants are recommended to interact with the mock-up Sina Weibo to get familiarize with the interface.

5. Participants are instructed to post a few non-sensitive posts to the mock-up Sina Weibo. If the participants have troubles constructing posts, the researchers will provide a list of posts that have neutral content to the participants to post to the mock-up Sina Weibo.

6. Participants are given a list of posts that contain censored keywords and are likely to be censored on Sina Weibo.

7. Participants are instructed to post some of these posts to the mock-up Sina

Weibo with the CENSE system disabled and then with the CENSE system enabled.

8. Participants are asked to complete a survey to evaluate the usefulness and the effectiveness of the CENSE system at the end of the study.

Throughout the study, the researchers will take note of the thoughts and comments that the participants made throughout the study. At the end of the study, the researchers will be available to explain the mechanics of the CENSE system if the participants have any questions.

By conducting the proposed think-aloud lab study, the usefulness and the effectiveness of the CENSE system can be evaluated by Chinese social media users, the targeted user group of the system, without compromising the privacy and security of the participants.

5.8 Circumventing Governmental Attacks

It is undeniable that systems that defy governmental control of the Internet such as CENSE will be subjected to network surveillance by the government. Moreover, in the case of CENSE, the Chinese government can easily render the system useless by simply banning the access to the backend server of the system. In this section, I explain how the CENSE system is designed to address these two issues.

5.8.1 Defense Against Network Tracing

The communication between the front-end client and the back-end server of CENSE system leaves a vulnerable point for government surveillance of users of the system. In this case, the government can monitor data sent over the Internet to identify the identity of CENSE users. In addition, the government can perform the man-in-the-middle attack to trick the users through the front-end client that they receive the data from the back-end server, but the users actually receive the data from the government.

CENSE is designed to take these possible attacks into account. To get rid of a possible vulnerable point in the communication between the front-end client and the back-end server, CENSE users can opt to run their own back-end servers on their local networks or even on the same machine as the front-end client. This way, the front-end client and the back-end server do not need to connect to the Internet in order to communicate with each other. Therefore, the users who use this set up of the CENSE system can ensure that their use of the system cannot be traced via the Internet.

5.8.2 Defense Against Access Restrictions

As explained above, the Chinese government can disable the functionality of CENSE by simply banning access to the backend servers of the system. In this case, the solution to defend against network tracing described in the previous section will also solve the problem of access ban to the backend servers. Because users deploy the backend servers on their own local network, it is impossible for the government to block access within users' own local network. Nevertheless, there is still one lingering problem related to the access ban by the government.

Because the CENSE system relies on two datasets of Chinese social media posts to properly provide the up-to-date list of censored keywords on Chinese social media to users. The first dataset is the uncensored posts on Chinese social media which is easily obtained in China. On the other hand, the second dataset—the censored posts on Chinese social media—is hard to obtain within China as social media curator sites such as Weiboscope and Freeweibo are sometimes blocked by the Great Firewall. Without the second dataset, CENSE will not be able to generate a list of censored keywords on Chinese social media for the users.

To get around this problem, CENSE can be periodically released to include an updated list of censored keywords at the time of release. For example, the system

can release a updated version every three months which includes the list of censored keywords at the time of release. This way, CENSE users who do not have access to censored posts on Chinese social media can still use the system with a list of censored keywords that is "semi-up-to-date".

5.9 Limitations and Future Work

The primary limitation encountered in my development of the CENSE system, which involves the sensitive topic of Internet censorship in China, was that an evaluation involving actual users could not be performed as an ethical research practice for reasons discussed in the previous section. Therefore, the system could not be evaluated via the real-world use case scenarios. However, I believe that the novelty of this system provides significant contributions to the field of Social Computing research, and users of Chinese social media will benefit from its development.

As future work, the CENSE system can be improved in several ways to increase its performance and adaptability to Chinese social media. First, as we saw previously, performance of the Censored Keyword Extractor and of the Homophone Transformer modules were not as efficient as they could be. Currently, execution time of both modules takes approximately one hour. With better optimization of the TF-IDF algorithm and the introduction of distributed computing algorithms such as Map Reduce, the performance of these two modules could be improved.

Second, the Homophone Transformer module currently relies on Da's character frequency list of Classical and Modern Chinese [25] to ensure that the generated homophones use Chinese characters having high frequency in Chinese language texts, resulting in high reader recognition and high entropy of generated words. However, Da's character frequency list is not updated with the language used on Chinese social media, which differ significantly from the language used in Chinese literature. Therefore, the system can improve on this aspect by generating its own character

frequency list based on Chinese social media data obtained from the Social Media Stream Watcher module. In this way, the generated homophones can be selected to better match the language used on Chinese social media, improve reader recognition, and generate increased entropy by matching with more false-positive words on social media.

To evaluate the CENSE system, I propose an evaluation plan through a lab study to evaluate the system's usefulness and effectiveness from the perspective of the users. While the proposed lab study in the previous section does not replace a deployment study, the lab study provides an opportunity for the system to be evaluated by the targeted user group.

My development of CENSE serves as a proof-of-concept that a real-time, automated Chinese social media censorship circumvention tool is viable and would be welcomed by Chinese social media users. I hope that the CENSE system will inspire researchers and designers to create tools to help social media users in China and other countries with repressive governments to be able to freely express their thoughts and opinions on social media.

CHAPTER VI

CONCLUSION

User participation is crucial to the longevity of online communities and social media platforms. The research literature has suggested that imposition of Internet censorship by repressive governments discourages their citizens from contributing user-generated content to the local platforms [95]. However, this is apparently not the case with China. There, although access to Western social media platforms such as Facebook and Twitter is prohibited, local versions such as Sina Weibo and WeChat flourish in spite of being subject to tight governmental control. In spite of censorship, there seems to be no lack of content and user participation on Chinese social media.

The work in this thesis revealed that the effects of censorship on Chinese social media users are predominantly *off-platform*. In other words, censorship has made users cautious about the type of content they post on social media and has therefore suppressed sensitive speech. While *on-platform* effects are detectable on users' social media profiles, these tend to wear off over time.

The primary reason that off-platform censorship effects are more apparent than on-platform ones is that Chinese social media users lack specific knowledge of what content is being censored. In my interview study, Chinese social media users stated that they self-censor out of caution whenever they think their posts could be considered "sensitive." Therefore, as a social computing researcher, I saw an opportunity to develop a tool to aid Chinese social media users in better understanding the censorship mechanism and thus enable them to experience increased freedom of expression by circumventing censorship.

To circumvent censorship, I developed a non-deterministic algorithm utilizing the

term frequency-inverse document frequency (TF-IDF) to generate homophones of Chinese words. These homophones, created from commonly used Chinese characters drawn from classical and modern Chinese, have the same sounds as the corresponding words the homophones are designed to replace and so trick censorship algorithms. In my experiments, replacing censored keywords in social media posts with their homophones extended the posts' lifetimes on Sina Weibo by three times compared to the original posts. Moreover, these transformed posts were readily understandable by Chinese native speakers. Based on its success, I employed this homophone transformation algorithm in the design of a real-time system, CENSE, which can not only enable Chinese social media users to circumvent censorship on social media but also allow them to better understand their censorship situations.

CENSE is composed of two primary components: a back-end server and a front-end client. The back-end server handles the logical operations needed to generate homophones for censored keywords, while the front-end client warns users of censorship and communicates suggested homophones substitutions. The back-end server first collects censored and uncensored posts from Sina Weibo, the Chinese version of Twitter, and applies the TF-IDF algorithm on these two datasets to identify censored keywords. The homophone-generation algorithm then generates using my homophone replacements for these keywords. The front-end client, a Google Chrome extension, automatically notifies users when their social media posts contain one or more censored keywords and, in the presence of censored keywords, suggests homophone replacements. Thus, the system not only gives users the option of neutralizing the effects of censorship but also allows them to see what specific words and topics are considered sensitive by their government. CENSE is one of the first automated systems to both bring transparency to the censorship process and allow users to circumvent censorship on Chinese social media.

6.1 Future Research Directions Stemming from this Work

The research in thesis has opened up several future research avenues in the fields of Social Computing and Social Systems research. In this section, I discuss each of these and suggest possible issues that can be explored in the future.

6.1.1 Adversarial Social Computing

The research literature in folk theories of social systems suggested that users of social systems often develop their own theories of how complex personalization algorithms that are driving these systems work [32]. On a similar note, Chinese social media users also have their speculations of how the censorship mechanism works as reflected in the results of the interview study presented in Chapter 3. However, there are two main differences between a social system's personalization algorithm and a censorship mechanism.

1. The censorship adversaries do not want users to understand the censorship mechanism. Although social systems such as Facebook and Youtube have not disclosed how exactly their personalization and recommendation algorithms work, they are up front with how user information are factored into the algorithms [22, 34]. On the other hand, Chinese social media platforms and the Chinese government have never released any information regarding the censorship mechanism; the only information available are uncovered by researchers such as [8, 60, 61, 124].

2. Users do not have influence over the censorship mechanism. As mentioned earlier, user information and actions on social systems often influence user experience on the systems through personalization algorithms. However, the censorship mechanism is only affected by current political situations.

I would like to advocate for a branch of Social Computing research that specifically

concerns issues arising from adversarial social systems. There is a need to gain a better understanding of how users on current social systems with underlying adversaries are behaving. In contrast with western social systems, users on adversarial social systems could develop *off-platform* behaviors, affecting their online experiences as demonstrated by Chinese social media users through my interview study in Chapter 3.

6.1.2 Systems Supporting Adversarial Social Media

I began this thesis by asking if the censorship apparatus deployed against sites like Sina Weibo exerts any user-level costs. It seems intuitive to us in western society, where free speech is a guarantee that the sporadic enactment of censorship might drive social media users underground, making them more likely to abandon their accounts or simply post less frequently. Indeed, if this were the case, one could use it as ammunition against other countries that might look to borrow the Chinese model. In other words: "It may work in the short term, but it ends up undermining your platform in the long term."

However, this prediction has not turned out to be true with respect to China, as was shown in the results presented in Chapter 3. Instead, on-platform effects of censorship appear to be limited, and the user base seems to have largely internalized caution around sensitive topics. In other words, censorship seems to work exactly as you would expect the state wants it to work. As a free speech advocate myself, this gives me pause. Perhaps there are other deterrents to deploying censorship, but decreased participation does not seem to be one of them. I worry that, as a consequence, other countries will deploy similar technologies in the future. The findings from my research encourage social systems researchers to create technologies that aid users in understanding, and potentially circumnavigating, elaborate censorship mechanisms, as I attempted to show in Chapter 4 and 5.

6.2 Summary of Findings

To summarize, the major findings of this thesis are as follows:

1. The user-level effects of censorship on Chinese social media are intrinsic rather than extrinsic. Users do not fully understand the censorship mechanism and deliberately self-censor out of caution. While the user-level effects of censorship on Chinese social media can be detected on platforms, they tend to wear out over time.

2. It is possible to circumvent censorship on Chinese social media by substituting censored keywords with computationally generated homophones. The lifetime of sensitive posts in which corresponding homophones replace censored keywords is three times longer than their original, unaltered counterparts. Moreover, Chinese native speakers can still fully understand the content of the transformed posts.

3. An automated, real-time system to circumvent censorship on Chinese social media is possible. In this particular case, I used a homophone generation algorithm I designed along with a censored keywords detection algorithm to provide users with current information on censorship on Chines social media. Together with a front-end client, this system could not only help Chinese social media users better understand the censorship mechanism but also circumvent it. Chinese social media users welcomed this idea of a tool that would help them circumvent censorship on social media. Although they did not see immediate use for such a tool, they stated that having access to it would be useful.

6.3 Concluding Remarks

The ideology under which this thesis was conducted views repression of political speech on social media as evil. It is true that not all content on the Internet should

be published for everyone to see. For example, a selfie promoting unhealthy behavior such as self-harm or an eating disorder should be moderated to remove the a bad influence it has on vulnerable population. However, I believe that the Internet can and should be a medium for healthy conversations on almost any topics. Unfortunately, this is not the case in China.

Understandably, the Chinese government seeks to control the media to avoid a repeat of the devastating events of the 1989 Tiananmen Square student demonstration. At the same time, I would argue that the government is infringing too greatly on the rights of Chinese citizens to criticize government officials and government policy. Basically, the Chinese government is disallowing the key factor needed for a democratic government: its citizens' ability to oversee their government. While introducing democracy to China is not (yet) the most pressing issue for which to advocate, in several cases, suppressing citizens' speech has led to abuses of power by government officials and significant violations of human rights.

It is now up to us as researchers and technology developers to lend a hand in improving the lives of Chinese Internet users. After all, technology, computers, and the Internet were created to make people's lives better and increase the quality of life. I believe that, through this work done in this thesis, I have done my part in helping fellow Internet users, and my hope is that this work will inspire others to do the same.

REFERENCES

[1] Akyildiz, I. F. and Wang, X., "A survey on wireless mesh networks," *IEEE Communications Magazine*, vol. 43, pp. S23–S30, sep 2005.

[2] Akyildiz, I. F., Wang, X., and Wang, W., "Wireless mesh networks: a survey," *Computer networks*, vol. 47, no. 4, pp. 445–487, 2005.

[3] Al-Akkad, A., Ramirez, L., Denef, S., Boden, A., Wood, L., Büscher, M., and Zimmermann, A., ""Reconstructing Normality": The Use of Infrastructure Leftovers in Crisis Situations As Inspiration for the Design of Resilient Technology," OzCHI '13, (New York, NY, USA), pp. 457–466, ACM, 2013.

[4] Al-ani, B., Mark, G., Chung, J., and Jones, J., "The Egyptian Blogosphere: A Counter-narrative of the Revolution," in *Cscw 2012*, CSCW '12, (New York, NY, USA), pp. 17–26, ACM, 2012.

[5] Aryan, S., Aryan, H., and Halderman, J. A., "Internet Censorship in Iran: A First Look," *Free and Open Communications on the Internet, Washington, DC, USA*, 2013.

[6] Athens Wireless Metropolitan Network, "AWMN Portal - The Front Page," 2014.

[7] Attia, A. M., Aziz, N., Friedman, B., and Elhusseiny, M. F., "Commentary: The impact of social networking tools on political change in Egypt's "Revolution 2.0"," *Electronic Commerce Research and Applications*, vol. 10, no. 4, pp. 369–374, 2011.

[8] Bamman, D., O'Connor, B., and Smith, N., "Censorship and deletion practices in Chinese social media," *First Monday*, vol. 17, mar 2012.

[9] BBC, "China issues new internet rules that include jail time - BBC News," 2013.

[10] Bond, R. M., Fariss, C. J., Jones, J. J., Kramer, A. D. I., Marlow, C., Settle, J. E., and Fowler, J. H., "A 61-million-person experiment in social influence and political mobilization," *Nature*, vol. 489, pp. 295–298, sep 2012.

[11] Boyd, D. and Ellison, N. B., "Social Network Sites: Definition, History, and Scholarship," *Journal of Computer-Mediated Communication*, vol. 13, no. 1, pp. 210–230, 2007.

[12] Braun, V. and Clarke, V., "Using thematic analysis in psychology," *Qualitative Research in Psychology*, vol. 3, pp. 77–101, jan 2006.

[13] Brown, I., "Internet Filtering - Be Careful What You Ask for," *Freedom and Prejudice: Approaches to Media and Culture*, pp. 74–91, jan 2007.

[14] Burke, M., Kraut, R., and Marlow, C., "Social capital on Facebook: Differentiating uses and users," in *CHI 2011, May 7–12, 2011, Vancouver, BC, Canada.*, pp. 571–580, 2011.

[15] Burnett, S., Feamster, N., and Vempala, S., "Chipping Away at Censorship Firewalls with User-Generated Content.," *USENIX Security Symposium*, 2010.

[16] Burnett, S. and Feamster, N., "Encore," *ACM SIGCOMM Computer Communication Review*, vol. 45, pp. 653–667, aug 2015.

[17] Campbell, C., "Internet Censorship Is Taking Root in Southeast Asia," 2013.

[18] Chen, L., Zhang, C., and Wilson, C., "Tweeting Under Pressure: Analyzing Trending Topics and Evolving Word Choice on Sina Weibo," COSN '13, (New York, NY, USA), pp. 89–100, ACM, 2013.

[19] Clayton, R., Murdoch, S. J., and Watson, R. N. M., "Ignoring the Great Firewall of China," in *Privacy Enhancing Technologies* (Danezis, G. and Golle, P., eds.), Lecture Notes in Computer Science, pp. 20–35, Springer Berlin Heidelberg, jan 2006.

[20] Cogburn, D. L. and Espinoza-Vasquez, F. K., "From Networked Nominee to Networked Nation: Examining the Impact of Web 2.0 and Social Media on Political Participation and Civic Engagement in the 2008 Obama Campaign," *Journal of Political Marketing*, vol. 10, no. 1-2, pp. 189–213, 2011.

[21] Cohen, R., "The Obama Connection," 2008.

[22] Covington, P., Adams, J., and Sargin, E., "Deep Neural Networks for YouTube Recommendations," in *Proceedings of the 10th ACM Conference on Recommender Systems - RecSys '16*, (New York, New York, USA), pp. 191–198, ACM Press, 2016.

[23] Crandall, J. R., Zinn, D., Byrd, M., and East, R., "ConceptDoppler: a weather tracker for internet censorship," in *Proceedings of the 14th ACM conference on Computer and communications security - CCS '07*, (Alexandria, VA), pp. 352–365, ACM Press, 2007.

[24] Custer, C., "China will embed police in internet company offices. But what does that mean?," 2015.

[25] Da, J., "Combined character frequency list of Classical and Modern Chinese," dec 2005.

[26] Dehejia, R. H. and Wahba, S., "Propensity Score-Matching Methods for Nonexperimental Causal Studies," *Review of Economics and Statistics*, vol. 84, pp. 151–161, feb 2002.

[27] Deibert, R. J. and Villeneuve, N., "Firewalls and power: An overview of global state censorship of the Internet," *Human rights in the digital age. London: GlassHouse*, 2004.

[28] Dibbell, J., "The Shadow Web," *Scientific American*, vol. 306, pp. 60–65, mar 2012.

[29] Dye, M., Anton, A., and Bruckman, A. S., "Early Adopters of the Internet and Social Media in Cuba," in *CSCW 2016*, (San Francisco, CA), 2016.

[30] Eriksson, J. and Giacomello, G., "Who controls what, and under what conditions," *International Studies Review*, 2009.

[31] Esarey, A. and Qiang, X., "Political Expression in the Chinese Blogosphere: Below the Radar," *Asian Survey*, vol. 48, pp. 752–772, oct 2008.

[32] Eslami, M., Karahalios, K., Sandvig, C., Vaccaro, K., Rickman, A., Hamilton, K., and Kirlik, A., "First I "like" it, then I hide it," in *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems - CHI '16*, (New York, New York, USA), pp. 2371–2382, ACM Press, 2016.

[33] Facebook, "Statement of Rights and Responsibilities," 2015.

[34] Facebook Help Center, "How News Feed Works."

[35] Feamster, N., Balazinska, M., Harfst, G., Balakrishnan, H., and Karger, D., "Infranet: Circumventing Web Censorship and Surveillance.," in *USENIX Security Symposium*, pp. 247–262, 2002.

[36] Filastò, A. and Appelbaum, J., "OONI: Open Observatory of Network Interference," in *Free and Open Communications on the Internet*, 2012.

[37] Freedman, J. L. and Fraser, S. C., "Compliance without pressure: The foot-in-the-door technique," *Journal of Personality and Social Psychology*, vol. 4, no. 2, pp. 195–202, 1966.

[38] Fu, K.-w., Chan, C.-h., and Chau, M., "Assessing Censorship on Microblogs in China: Discriminatory Keyword Analysis and the Real-Name Registration Policy," *IEEE Internet Computing*, vol. 17, pp. 42–50, may 2013.

[39] Gilbert, E., "Open Book: A Socially-inspired Cloaking Technique that Uses Lexical Abstraction to Transform Messages," in *Proceedings of the ACM CHI'15 Conference on Human Factors in Computing Systems*, vol. 1, pp. 477–486, 2015.

[40] Gong, H. and Yang, X., "Digitized parody: The politics of egao in contemporary China," *China Information*, vol. 24, pp. 3–26, mar 2010.

[41] Guifi.net, "guifi.net - Open, Libre and Neutral Telecommunications Network — guifi.net," 2014.

[42] Hassid, J., "Safety Valve or Pressure Cooker? Blogs in Chinese Political Life," *Journal of Communication*, vol. 62, pp. 212–230, apr 2012.

[43] Hiruncharoenvate, C., Lin, Z., and Gilbert, E., "Algorithmically Bypassing Censorship on Sina Weibo with Nondeterministic Homophone Substitutions," in *International AAAI Conference on Web and Social Media*, 2015.

[44] Ho, D. E., Imai, K., King, G., and Stuart, E. A., "MatchIt: Nonparametric Preprocessing for Parametric Causal Inference," *Journal of Statistical Software*, vol. 42, no. 8, 2011.

[45] Hourcade, J. P. and Bullock-Rest, N. E., "HCI for Peace: A Call for Constructive Action," in *CHI 2011*, CHI '11, (New York, NY, USA), pp. 443–452, ACM, 2011.

[46] Howard, P. N. and Hussain, M. M., "The Role of Digital Media," *Journal of Democracy*, vol. 22, no. 3, pp. 35–48, 2011.

[47] Huang, H., Wen, Z., Yu, D., Ji, H., Sun, Y., Han, J., and Li, H., "Resolving Entity Morphs in Censored Data.," pp. 1083–1093, 2013.

[48] Hutchins, E., "How a cockpit remembers its speeds," *Cognitive Science*, vol. 19, pp. 265–288, jul 1995.

[49] Iclab, "Centinel," 2016.

[50] Incitez China, "481.5 Million Social Media Users in China in 2015," 2016.

[51] Instagram, "Terms of Use," 2013.

[52] Internet Live Stats, "Internet Users by Country (2016)," 2016.

[53] Iqbal, M., Wang, X., Wertheim, D., and Zhou, X., "SwanMesh: A Multicast Enabled Dual-Radio Wireless Mesh Network for Emergency and Disaster Recovery Services," *Journal of Communications*, vol. 4, jun 2009.

[54] Jones, B., Lee, T.-W., Feamster, N., and Gill, P., "Automated Detection and Fingerprinting of Censorship Block Pages," in *Proceedings of the 2014 Conference on Internet Measurement Conference*, (New York, New York, USA), pp. 299–304, ACM Press, nov 2014.

[55] Kalathil, S. and Boas, T. C., "The Internet and state control in authoritarian regimes: China, Cuba and the counterrevolution," *First Monday*, vol. 6, aug 2001.

[56] Kazansky, B., "In Red Hook, Mesh Network Connects Sandy Survivors Still Without Power," nov 2012.

[57] Kessler, S., "Why Social Media Is Reinventing Activism," oct 2010.

[58] Kessler, S., "How Occupy Wall Street Is Building Its Own Internet [VIDEO]," 2011.

[59] King, G. and Nielsen, R., "Why propensity scores should not be used for matching," 2016.

[60] King, G., Pan, J., and Roberts, M. E., "How Censorship in China Allows Government Criticism but Silences Collective Expression," *American Political Science Review*, vol. 107, pp. 326–343, may 2013.

[61] King, G., Pan, J., and Roberts, M. E., "Reverse-engineering censorship in China: Randomized experimentation and participant observation," *Science*, vol. 345, p. 1251722, aug 2014.

[62] Kloc, J., "Greek community creates an off-the-grid Internet," aug 2013.

[63] Koku, E., Nazer, N., and Wellman, B., "Netting Scholars Online and Offline," *American Behavioral Scientist*, vol. 44, pp. 1752–1774, jun 2001.

[64] Kraut, R., Patterson, M., Lundmark, V., Kiesler, S., Mukopadhyay, T., Scherlis, W., Mukophadhyay, T., and Scherlis, W., "Internet paradox: A Social Technology That Reduces Social Involvement and Psychological Well-Being?," *American Psychologist*, vol. 53, no. 9, pp. 1017–1031, 1998.

[65] Lampe, C., Ellison, N., and Steinfield, C., "A Face(book) in the Crowd: Social Searching vs. Social Browsing," *Proceedings of the 2006 20th Anniversary Conference on Computer-Supported Cooperative Work CSCW '06*, pp. 167–170, 2006.

[66] Leberknight, C. S., Poor, H. V., Chiang, M., and Wong, F., "A taxonomy of Internet censorship and anti-censorship," *Fifth International Conference on Fun with Algorithms*, 2010.

[67] Leibold, J., "Blogging Alone: China, the Internet, and the Democratic Illusion?," *The Journal of Asian Studies*, vol. 70, pp. 1023–1041, oct 2011.

[68] Ling, W. and Xiang, G., "Sina Weibo API Guide."

[69] Lu, J. and Qiu, Y., "Microblogging and Social Change in China," jul 2013.

[70] MacKinnon, R., "Flatter world and thicker walls? Blogs, censorship and civic discourse in China," *Public Choice*, vol. 134, no. 1-2, pp. 31–46, 2008.

[71] McGregor, S. E., "Can mesh networks and offline wireless move from protest tools to news?," 2014.

[72] McLeod, J. M., Daily, K., Guo, Z., Eveland, W. P., Bayer, J., Yang, S., and Wang, H., "Community Integration, Local Media Use, and Democratic Processes," *Communication Research*, vol. 23, no. 2, pp. 179–209, 1996.

[73] McLeod, Mira Sotirovic, J. M., "Values, Communication Behavior, and Political Participation," *Political Communication*, vol. 18, no. 3, pp. 273–300, 2001.

[74] Murdoch, S. J. and Anderson, R., "Tools and technology of Internet filtering," *Access Denied: The Practice and Policy of Global Internet Filtering, ed. Ronald J. Deibert, John Palfrey, Rafal Rohozinski, and Jonathan Zittrain (Cambridge, MA: MIT Press, 2008)*, 2008.

[75] Nabi, Z., "The Anatomy of Web Censorship in Pakistan," jul 2013.

[76] Nasr, L., "The Case of the Disappearing Activists: The Fight for Freedom of Speech in China — LSE Human Rights," 2016.

[77] Norris, P., "Does Television Erode Social Capital? A Reply to Putnam," *American Political Science Association*, vol. 29, no. 3, pp. 474–480, 1996.

[78] OpenNet Initiative, "OpenNet Initiative," 2016.

[79] Pei, M., "China's tragic crackdown on social media activism," 2013.

[80] Poetranto, I., "Toward an open, free and secure Internet," 2014.

[81] Putnam, R. D., "Tuning in, tuning out: The strange disappearance of social capital in America," *Political science and politics*, vol. 28, no. 4, p. 664, 1995.

[82] Putnam, R. D., "Bowling alone: America's declining social capital," *Journal of democracy*, vol. 6, no. 1, pp. 65–78, 1995.

[83] Putnam, R. D., *Bowling alone: The collapse and revival of American community*. 2000.

[84] Qiang, X., "The 'blog' revolution sweeps across China," 2004.

[85] Quinn, M. J., *Ethics for the Information Age*. 6 ed., 2014.

[86] Rao, J. R., "Can Pseudonymity Really Guarantee Privacy?," in *Proceedings of the 9th USENIX Security Symposium*, no. August, 2000.

[87] Reuters, "Phone companies in Brazil blocking YouTube," 2007.

[88] Roberts, M. E., "Experiencing Censorship Emboldens Internet Users and Decreases Government Support in China," *Unpublished Working Paper*, 2015.

[89] Rosenbaum, P. R. and Rubin, D. B., "Constructing a Control Group Using Multivariate Matched Sampling Methods That Incorporate the Propensity Score," *The American Statistician*, vol. 39, p. 33, feb 1985.

[90] Sanchez, L., "Commentary: GOP needs to catch up to Obama's Web savvy," 2008.

[91] Sfakianakis, A., Athanasopoulos, E., and Ioannidis, S., "CensMon: A Web Censorship Monitor," *USENIX Workshop on Free and Open Communications on the Internet*, 2011.

[92] Shah, D., Schmierbach, M., Hawkins, J., Espino, R., and Donavan, J., "Nonrecursive Models of Internet Use and Community Engagement: Questioning Whether Time Spent Online Erodes Social Capital," *Journalism & Mass Communication Quarterly*, vol. 79, no. 4, pp. 964–987, 2002.

[93] Shah, D. V., Kwak, N., and Holbert, R. L., ""Connecting" and "Disconnecting" With Civic Life: Patterns of Internet Use and the Production of Social Capital," *Political Communication*, vol. 18, no. 2, pp. 141–162, 2001.

[94] Shirky, C., "Political Power of Social Media-Technology, the Public Sphere Sphere, and Political Change, The," *Foreign Aff.*, vol. 90, p. 28, 2011.

[95] Shklovski, I. and Kotamraju, N., "Online Contribution Practices in Countries That Engage in Internet Blocking and Censorship," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '11, (New York, NY, USA), pp. 1109–1118, ACM, 2011.

[96] Staff, "Social media – tool of revolution or repression?," jan 2011.

[97] Statista, "VPN usage in selected countries 2014," 2014.

[98] Statista, "VPN: penetration rate in selected countries 2015," 2015.

[99] Stern, R. E. and Hassid, J., "Amplifying Silence: Uncertainty and Control Parables in Contemporary China," *Comparative Political Studies*, vol. 45, no. 10, pp. 1230–1254, 2012.

[100] Stevenson, C., "Breaching the Great Firewall: China's Internet Censorship and the Quest for Freedom of Expression in a Connected World," *Boston College International and Comparative Law Review*, vol. 30, 2007.

[101] Stewart, F., Holdstock, D., and Jarquin, A., "Root causes of violent conflict in developing countriesCommentary: Conflict—from causes to prevention?," *BMJ*, vol. 324, pp. 342–345, feb 2002.

[102] Sun, J., "jieba: Chinese Words Segmentation Utilities," 2015.

[103] Taubman, G. L., "Non-democratic legitimacy and access to the Web," sep 2002.

[104] The Guardian, "Chinese police arrest 15,000 for cybercrimes," 2015.

[105] Thompson, C., "How to Keep the NSA Out of Your Computer," sep 2013.

[106] Thompson, D., "Why the Internet Is About to Replace TV as the Most Important Source of News," 2012.

[107] Tufekci, Z. and Wilson, C., "Social media and the decision to participate in political protest: Observations from Tahrir Square," *Journal of Communication*, vol. 62, no. 2, pp. 363–379, 2012.

[108] Twitter, "Removal requests — Transparency report," 2015.

[109] Twitter, "Twitter Terms of Service," 2016.

[110] VPN Gate, "VPN Gate Overview," 2014.

[111] Wang, A., "Weibo Search Users Insights in 2015," 2015.

[112] Wang, D. and Mark, G., "Internet Censorship in China," *ACM Transactions on Computer-Human Interaction*, vol. 22, pp. 1–22, nov 2015.

[113] Warf, B., "Geographies of global Internet censorship," *GeoJournal*, vol. 76, pp. 1–23, nov 2010.

[114] Wellman, B., Haase, A. Q., Witte, J., and Hampton, K., "Does the Internet Increase, Decrease, or Supplement Social Capital? Social Networks, Participation, and Community Commitment," *American Behavioral Scientist*, vol. 45, pp. 436–455, nov 2001.

[115] Wikipedia, "Cute cat theory of digital activism," may 2014.

[116] Wikipedia, "List of websites blocked in China," jul 2014.

[117] Wikipedia, "Microblogging in China," jan 2015.

[118] Wikipedia, "Countries blocking access to The Pirate Bay," 2016.

[119] Wu, F. and Yang, S., "Web 2.0 and Political Engagement in China," *VOLUNTAS: International Journal of Voluntary and Nonprofit Organizations*, jul 2015.

[120] Wulf, V., Misaki, K., Atam, M., Randall, D., and Rohde, M., "'On the Ground' in Sidi Bouzid: Investigating Social Media Use During the Tunisian Revolution," in *CSCW 2013*, CSCW '13, (New York, NY, USA), pp. 1409–1418, ACM, 2013.

[121] Yuan, L., "China enlists the public in its ongoing campaign to censor the Internet," 2001.

[122] Zhang, B., Huang, H., Pan, X., Ji, H., Knight, K., Wen, Z., Sun, Y., Han, J., and Yener, B., "Be Appropriate and Funny: Automatic Entity Morph Encoding," 2014.

[123] Zhang, W., Johnson, T. J., Seltzer, T., and Bichard, S. L., "The Revolution Will Be Networked: The Influence of Social Networking Sites on Political Attitudes and Behavior," *Social Science Computer Review*, jun 2009.

[124] Zhu, T., Phipps, D., Pridgen, A., Crandall, J. R., and Wallach, D. S., "The velocity of censorship: High-fidelity detection of microblog post deletions," *arXiv preprint arXiv:1303.0597*, 2013.

[125] Zuckerman, E., "The Cute Cat Theory Talk at ETech," mar 2008.

[126] Zuckerman, E., Roberts, H., and Palfrey, J., "2007 Circumvention Landscape Report: Methods, Uses, and Tools," no. March, pp. 1–95, 2009.