Algorithmically Bypassing Censorship on Sina Weibo with Nondeterministic Homophone Substitutions

Chaya Hiruncharoenvate, Zhiyuan Lin and Eric Gilbert
School of Interactive Computing & GVU Center
Georgia Institute of Technology
{chaya, zlin48}@gatech.edu, gilbert@cc.gatech.edu

Abstract

Like traditional media, social media in China is subject to censorship. However, in limited cases, activists have employed *homophones* of censored keywords to avoid detection by keyword matching algorithms. In this paper, we show that it is possible to scale this idea up in ways that make it difficult to defend against. Specifically, we present a non-deterministic algorithm for generating homophones that create large numbers of false positives for censors, making it difficult to locate banned conversations. In two experiments, we show that 1) homophone-transformed weibos posted to Sina Weibo remain on-site three times longer than their previously censored counterparts, and 2) native Chinese speakers can recover the original intent behind the homophone-transformed messages, with 99% of our posts understood by the majority of our participants. Finally, we find that coping with homophone transformations is likely to cost the Sina Weibo censorship apparatus an additional 15 hours of human labor per day, per censored keyword. To conclude, we reflect briefly on the opportunities presented by this algorithm to build interactive, client-side tools that promote free speech.

Introduction

Social media sites have become powerful tools for citizens and activists. As demonstrated by recent uprisings across the world—from Tunisia to Egypt to Turkey—social media can play a central role in organizing citizens to collectively act against repressive regimes (Al-Ani et al. 2012; Wulf et al. 2013). Zuckerman explains this phenomenon with his "Cute Cat Theory of Digital Activism," whereby activists use general-purpose tools (e.g, Facebook and Twitter) instead of creating their own platforms to avoid getting their dedicated platforms shut down by governments (Zuckerman 2008).

The situation is very different in China. There, only local replicas of social media—such as the largest Chinese social networking site Sina Weibo, which effectively emulates Twitter—are allowed inside the country so that they can be closely monitored (and if need be, censored) by government officials. Until recently, we did not understand in detail how the censorship apparatus works on sites like Sina Weibo. However, in a recent paper, King and colleagues (2014) reverse-engineered the mechanics of censorship on Sina Weibo. To do this, King et al. 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 massive numbers of human censors also take part in the process. As it is central to the work presented here, we reproduce their main findings in Figure 2.

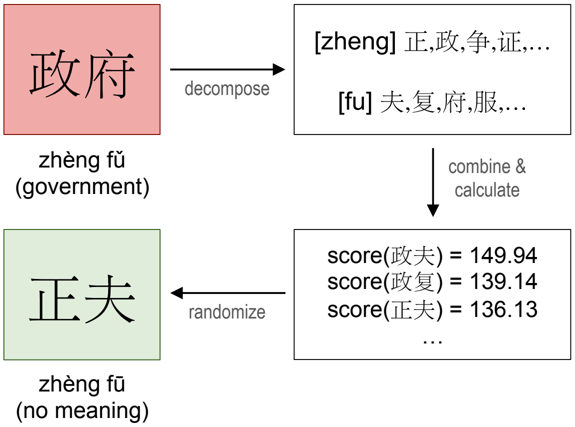

Figure 1: A high-level overview of our homophone generation algorithm. The original censored term, 政府, translated as "government," undergoes a process where constituent sounds generate a large number of possible character substitutions, those substitutions are combined, and then scored for their ability to create confusion for censors.

However, at least in some cases, Sina Weibo users have found a way around this censorship apparatus: homophones. Unlike English, certain properties of the Chinese language make it easy to construct words that sound nearly identical to other words, yet have completely different meanings. For instance, when a river crab meme spread across Sina Weibo, it did not really refer to river crabs. Rather, it stood for a protest against Internet censorship, as the word for *harmonize* (和谐, pronounced *hé xié*,), slang for *censorship*, is a homophone of the word for *river crab* (河蟹, pronounced *hé xiè*) (Zuckerman 2008).

In this paper, we show that it is possible to scale this idea up by automatically computing homophones. Moreover, these homophones can be computed in such a way as to present problems for the system depicted by King et al. in Figure 2. Specifically, we present a non-deterministic algorithm that generates homophones employing high-frequency characters, which in turn generates large numbers of false positives for Sina Weibo censors (see Figure 1). We also present the results of two experiments where we use this homophone algorithm to transform weibos (posts on Sina Weibo) known to have been previously censored. In the first experiment, we posted the homophone-transformed weibos to Sina Weibo and found that while both previously censored and homophone-transformed weibos ultimately got censored at the same rate, homophone-transformed weibos lasted on the site *three times as long* as their counterparts. In the second experiment, we show that native Chinese speakers on Amazon Mechanical Turk can understand these homophone-transformed weibos, with 99% of our posts clearly understood by the majority of our workers.

Finally, we believe it would cost the Sina Weibo censorship apparatus significant time and human resources to defend against this approach. Via an empirical analysis, we find that adversaries cannot simply add all homophones of censored keywords to a blocked keyword list because it would mistakenly censor a large portion of Sina Weibo's daily messages (one estimate in this paper suggests a figure of 20M posts per day, or 20% of daily messages). Rather, it seems likely that Sina Weibo would have to turn to human labor to defeat it. Based on previous Sina Weibo scholarship, we estimate that the technique proposed in this paper would cost site operators an additional *15 human-hours per day, per censored keyword*, a significant figure given that many thousands of banned keywords may be in place at any given time.

Related Work

Recent events have shown that the popularity and ubiquity of social media have created a new phenomenon where collective actions against powerful entities, such as repressive governments and political figures, have been organized and facilitated through social media (Al-Ani et al. 2012; Wulf et al. 2013). Zuckerman has called the phenomenon the "Cute Cat Theory of Digital Activism:" political activists blend in with normal Internet users on everyday platforms such as Facebook, Flickr, and Twitter (Zuckerman 2008). This allows activists to be more immune to government censorship because shutting down popular web services would provoke a larger public uproar than shutting down dedicated platforms for activism. However, the theory does not apply to the Chinese Internet, where popular social media sites such as Facebook and Twitter are completely blocked. Instead, local replicas such as Renren (a replica of Facebook) and Sina Weibo (a replica of Twitter) are deployed to support close content monitoring and censorship. Moreover, the Internet traffic in and out of the country has to pass through another layer of censorship in the form of government-mandated firewall which inspects every HTTP request for censored keywords (Clayton, Murdoch, and Watson 2006).

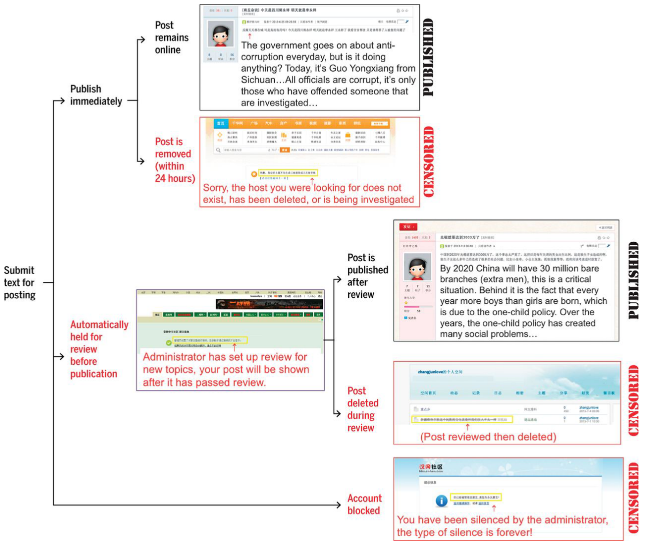
Submit text for posting
Publish immediately
Post remains online
The government goes on about anti-corruption everyday, but is it doing anything? Today, it's Guo Yongxiang from Sichuan…All officials are corrupt, it's only those who have offended someone that are investigated…
PUBLISHED
Post is removed (within 24 hours)
Sorry, the host you were looking for does not exist, has been deleted, or is being investigated
CENSORED
Automatically held for review before publication
Administrator has set up review for new topics, your post will be shown after it has passed review.
Post is published after review
By 2020 China will have 30 million bare branches (extra men), this is a critical situation. Behind it is the fact that every year more boys than girls are born, which is due to the one-child policy. Over the years, the one-child policy has created many social problems…
PUBLISHED
Post deleted during review
(Post reviewed then deleted)
CENSORED
Account blocked
You have been silenced by the administrator, the type of silence is forever!
CENSORED

Figure 2: Chinese censorship decision tree, reproduced from King et al. (2014).

Previous research on censorship in Chinese social media and blogs has surveyed how censorship is practiced and the range of content that is likely to be censored. MacKinnon tested the censorship practices of 15 blog service providers in China by posting controversial topics to these sites, finding that censorship practices were highly decentralized. These censorship practices varied based on several properties of the service providers—such as political views, sizes, public attention on front pages, and contacts they had with government units (MacKinnon 2009). Bamman et al. investigated censorship and message deletion practices on Sina Weibo. They found that there are some politically sensitive terms that lead to higher deletion rates as compared to a baseline. Furthermore, posts originating from regions in conflict, such as Tibet and Qinghai, were also deleted at a higher rate than posts from other areas of China (Bamman, O'Connor, and Smith 2012). King et al. added that posts that promote collective actions—regardless of their pro- or anti-government point of view—are mainly censored (King, Pan, and Roberts 2013). Their later work on establishing a social networking site in China produced a decision tree (Figure 2) that shows how and when posts are subject to censorship (King, Pan, and Roberts 2014).

Earlier scholars have found that Chinese Internet users were already using several properties of the Chinese language—such as decomposition of characters, translation, and creating nicknames—to circumvent adversaries, creating *morphs*, or aliases to hide the original words (Chen, Zhang, and Wilson 2013; Fu, Chan, and Chau 2013; Huang et al. 2013; Zhang et al. 2014; Zuckerman 2008). One of the best-known methods of creating morphs is using homophones. Homophones are common in Chinese due to the large number of characters with only a handful of corresponding sounds. 80% of the monosyllable sounds are ambiguous, with half of them having five or more corresponding characters (Li

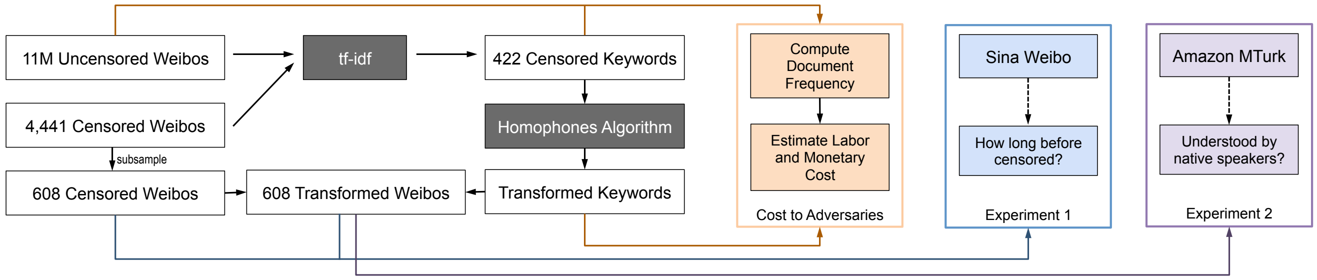

Figure 3: An overview of the datasets, methods, algorithms and experiments used in this work.

and Yip 1996). Zhang et al. gave examples of morphs created to circumvent censorship and compared human-generated and system-generated morphs. They found that while human-generated morphs were more appropriate and entertaining, they were easier to be discovered by an automatic morph decoder (Zhang et al. 2014). Huang et al. showed that it is possible to computationally resolve commonly-used morphs to their original forms when both censored and uncensored data are available (Huang et al. 2013). However, censorship adversaries can easily defend against commonly-used morphs by including them in the list of blocked keywords. Our goal is to create transformations of censored keywords such that the cost to adversaries outweighs the ambiguity created.

Countering Censorship with Homophones

Based on King et al.'s censorship decision tree (Figure 2), we speculated that it may be possible to consistently subvert Sina Weibo's censorship mechanisms by bypassing the initial review, thereby increasing the chance that posts will be published immediately. Both Bamman et al. (2012) and King et al. (2014) suggest that keyword detection plays a significant role in the censorship apparatus. The key insight of this paper is to computationally (and near optimally) alter the content of a post by *replacing censored keywords with homophones.* As this is already an emergent practice on Sina Weibo today, we expect that this transformation may allow native speakers to understand the original intent of the posts, given their awareness of the general topic of the posts (a claim we aim to experimentally verify). 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 Sina Weibo's current censorship apparatus.

In this paper, we explore these ideas in the form of three research questions bound together by the common theme of supporting free speech in Chinese social media:

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 3 presents an overview of the methods employed in this paper. It includes the two experiments that we performed to investigate our research questions, along with a cost analysis in which we estimate the costs site owners will have to pay in order to defend against our approach.

Datasets and Methods

In this paper, we often use the term *weibo* (微博)— which translates to "microblog"—when referring to social media posts in our dataset. The term *weibo* on Sina Weibo is roughly equivalent to *tweet* on Twitter.

To answer the research questions above, we first needed to identify censored keywords. We obtained two datasets to do so, comprising more than 11 million weibos. The first dataset consists of 4,441 weibos that are confirmed to be censored on Sina Weibo. We gathered this dataset from the website Freeweibo[1]; Freeweibo curates weibos from popular accounts on Sina Weibo. Similar to Weiboscope (Fu, Chan, and Chau 2013), Freeweibo also detects whether each weibo has been censored. Freeweibo displays the top 10 "hot search" keywords that were searched through their website at any unspecified time period. We obtained all hot search keywords that contain only Chinese 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, we assume as ground truth that the hot search keywords were the factor that led to censorship. We 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

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

not censored, people could simply do a search for them on Sina Weibo. In this manner, we collected a dataset of 4,441 censored weibos which were posted from October 2, 2009–November 20, 2014. Our two experiments on Sina Weibo itself and on Amazon Mechanical Turk rely on this dataset.

The second dataset consists of weibos from the public timeline of Sina Weibo. We used the Sina Weibo Open API to obtain these weibos available, again from October 13, 2014–November 20,2014, accumulating 11,712,617 weibos. We employ this corpus of weibos from the public timeline in our censored keyword extraction, homophone generation, and exploration of RQ3, the costs posed to the adversary in adapting to our homophone-generation technique.

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, we decided against using a pre-existing dictionary to extract words and phrases from our censored weibo dataset. Instead, we generated all two, three, and four-character words/phrases from the censored weibo dataset. We remove the terms that appear less than 10 times in the combined dataset of censored and uncensored weibos to ensure that the remaining terms commonly appear in social media. Then, we 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. We consider terms with tf-idf score in the top-decile to likely be censored keywords. We add to this computationally-inferred list the the hot search keywords from Freeweibo. In total, we therefore have 608 unique combinations of censored keywords. For each combination, we took the latest weibo in the censored dataset to form the small dataset of 608 weibos to explore in our experiments. (Our experimental methodologies, explained in greater detail later, carry a cost associated with each weibo in the dataset. We created a subsample for this reason.)

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 our work, we 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 our homophone generation procedure employs (to avoid generating very rare terms). We calculated the character frequency from our 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, we assign the frequency of the character to all sounds equally since we do not have information about the frequency distribution of the sounds. Then, for each of the 419 root sounds, we calculated the percentile of each character with that root sound based on its frequency.

Algorithm 1: Homophone generation

```
GetTopHphone
Input: $W$: Word for which to generate homophone
Output: $\widetilde{W}$: A homophone of $W$ with frequency score
        in the top k
$\widetilde{W_h} \leftarrow GenHphone(W)[rand(1, k)]$
$n \leftarrow len(W)$
if $n < 4$ then
    $\widetilde{W} \leftarrow \widetilde{W_h}$
else if $n = 4$ then
    $\widetilde{W} \leftarrow rand(\{\tilde{w}_h^1\tilde{w}_h^2, \tilde{w}_h^3\tilde{w}_h^4\})$
else if $n = 5$ then
    $\widetilde{W} \leftarrow rand(\{\tilde{w}_h^1\tilde{w}_h^2, \tilde{w}_h^3\tilde{w}_h^4\tilde{w}_h^5, \tilde{w}_h^1\tilde{w}_h^2\tilde{w}_h^3, \tilde{w}_h^4\tilde{w}_h^5\})$
return $\widetilde{W}$

GenHphone
Data: $C \leftarrow$ List of all characters in our frequency list
Input: $W \leftarrow$ Word for which to generate homophones
Output: $h_{topk} \leftarrow$ List of homophones with frequency
        score in the top k
for $w^i$ in $W$ do
    $h_i \leftarrow \{\tilde{w}^i : \tilde{w}^i \in C, sound(\tilde{w}^i) = sound(w^i)\}$
$h \leftarrow \{(\tilde{w} = \tilde{w}^1 \ldots \tilde{w}^n, score = \sum_i^n p(\tilde{w}^i)) : \tilde{w}^i \in h_i\}$
$h \leftarrow h - \{W\}$
$h_{topk} \leftarrow sortByScore(h, desc)[1 : k]$
return $h_{topk}$
```

To summarize, for a character c with corresponding sound r, we calculated its percentile p based on its frequency compared to other characters that also have the sound r. For each censored word W with characters $w^1w^2 \ldots w^n$, we can obtain its homophones $\widetilde{W_i}$ by combining the homophones of each character $\tilde{w}_i^1\tilde{w}_i^2 \ldots \tilde{w}_i^n$. Then, we use 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 1 shows an example of our algorithm generating a homophone for the censored keyword 政府 (government).

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

Because our algorithm ultimately interacts with censorship adversaries (something we describe in more detail in the *Cost to adversaries* section), we choose 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, we will show in Experiment 2 that native speakers can still infer the content of transformed weibos from shortened homophones. In our dataset, the maximum length of censored keywords is 5 characters. Therefore, we divide 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. Algorithm 1 summarizes this process in pseudocode.

Experiments

To address RQ1 and RQ2, we used an experimental approach. We took the 608 weibos from the subsampled dataset and transformed them by replacing their censored keywords with homophones generated from our non-deterministic algorithm presented above. We performed two experiments, each attempting to answer one of the research questions.

Experiment 1: Reposting to Sina Weibo. To answer RQ1, we posted the transformed content weibos to Sina Weibo using multiple newly created accounts. We measured the time it took for the weibos to get deleted or for the accounts to get banned. For comparison, we also posted originally censored (untransformed) weibos back to Sina Weibo and measured the same variables. We 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). We 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. (2014) method). Thus, the age of weibos has resolution at the minute timescale. The reason we needed a viewing account is that unregistered visitors can only view the first page of another user's timeline. In order to retrieve all of our posts, we 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 (King, Pan, and Roberts 2014; Zhu et al. 2013). Relying on this result, we monitor our posts from their posting time to 48 hours after they were posted.

Experiment 2: Amazon Mechanical Turk. To answer RQ2, we employed the online labor market Amazon Mechanical Turk (AMT) to hire native Chinese speakers to investigate if they could understand the weibos we transformed using homophones. We 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]." We then asked our participants 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 our tasks were native Chinese speakers, we provided the instructions and questions only in Chinese, accompanying it with an English message asking non-Chinese speakers not to complete the task. Each HIT (Human Intelligent Task) is comprised of four weibos (asking workers to answer a total of 12 questions.) We paid workers 20 cents for each HIT they completed. Workers were allowed to complete as many HITs as they wanted, up to 152 HITs (608 weibos.) For each HIT, we obtain completed work from 3 independent workers.

Results

Next, we 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).

Experiment 1: Censorship effects (RQ1)

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

- *Active* accounts can perform all activities on the site—logging in, posting, reading other users' timeline. Our 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 our 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

[2] English translation of original Chinese instructions.

	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

Table 1: Number of Weibo posts that survived through each stage of censorship.

blocked accounts, the same message was shown when our viewing accounts visited the timelines of *frozen* accounts.

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

For each originally censored weibo in our dataset, we posted it and its homophone-transformed version (totaling 1,216 weibos) back to Sina Weibo from our accounts. Throughout the rest of the paper, we refer to the posts we posted back to Sina Weibo as *original posts* and *transformed posts* based on their conditions. There are four progressive states that both types of our 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 our viewing accounts were able to see within 48 hours after they were posted.
- *Removed* posts are *published* posts that our 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 are not visible at the 48-hour mark for any reasons, including account termination.

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

Keyword transformations & censorship. Of the 1,216 weibos we 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 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 1 shows the number of weibo posts our 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.

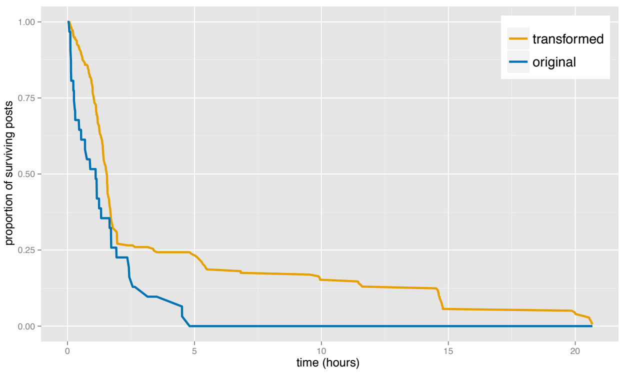

Figure 4: Proportion of *removed* posts surviving censorship, normalizing to posts' adjusted age. X-axis: Adjusted age; Y-axis: Proportion of *removed* posts.

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, we accounted for the variation in the distribution of the posted dates of censored weibos in our dataset by using the ratio of between the number removed posts (transformed and original) 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.

Experiment 2: Interpretability (RQ2)

In Experiment 2, 22 workers completed 456 assignments. Each assignment contains 4 different transformed weibos, resulting in 1,824 *impressions* of our 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 2 summarizes the statistics of weibos and worker impressions that reported confusion.

	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%)

Table 2: Number of impressions, weibos and workers' understanding of weibo content.

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 our 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, we consider an answer from our workers to be correct if either (1) it is the same as the keyword, (2) it is a substring of the keyword, or (3) the keyword is its substring. Then, we calculate 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 both transformed and original keywords, neither do they have significant effects on workers' understanding of the content of weibos.

In summary, we 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 our dataset, the majority of workers were able to understand the content from reading only the transformed posts.

Analysis: Cost to adversaries (RQ3)

Finally, we explore 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 paper, as well as what costs might be associated with those adaptations. As our 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 our 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 our approach. Next, we analyze 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 our dataset of 4,441 censored weibos, there were a total of 422 censored keywords, and our algorithm generated 8,400 unique homophones that have the frequency score in the top $k = 20$. We calculated the document frequency (one weibo treated as one document) of the homophones in our 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.) Our 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 $\widetilde{W_1}...\widetilde{W_k}$, we calculate 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, we consider the document frequency of all possible shortened keywords to be the number of false positive generated.

Then, for each censored keyword W, we calculate the average false positives generated over all of its homophones. We then calculate the average false positive generated in our dataset over all censored keywords. Algorithm 2 summarizes this process in pseudocode, the method used to calculate the number of false positive weibos for each censored keyword.

On average, each of our censored keywords matches 5,510 weibos in the uncensored corpus. Our uncensored sample corpus is only a fraction of the actual posts on Sina Weibo; there are approximately 100 million weibos made daily on

Algorithm 2: 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
          $S_i \leftarrow \{u \in U : u$ contains $\widetilde{W_i}\}$
      else
          $\widetilde{W_i'} \leftarrow \{$all shortened versions of $\widetilde{W_i}\}$
          $S_i \leftarrow \{u \in U : u$ contains any of $\widetilde{W_i'}\}$
  $k \leftarrow |\bigcup S_i|$
  return $k$
```

Sina Weibo (Zhu et al. 2013). Scaling the figure above to the actual amount of weibos sent daily, our 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 (Zhu et al. 2013), it would take more than 15 new human-hours each day to filter the false-positive weibos generated from each homophone-transformed keywords.

Discussion

First, we 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, we found that Turkers who natively speak Chinese can interpret the altered message. The datasets and methods used in this paper 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 we 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 we 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 our 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 our dataset discussed events that ended long before the time we 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, we measured the *relative* decrease in censorship after applying homophone transformations to our corpus.

Using homophones to transform censored keywords proved easy to understand by native speakers from the results of Experiment 2. None of our workers were confused with the content of 559 out of 608 (91.94%) transformed weibos, and the majority of our workers understood nearly all of our 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. Our 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.

Limitations

For practical and ethical reasons, we did not re-appropriate existing accounts for use in our experiments. They might be compromised and potentially even endanger the people operating them. Although the Real Name Policy is not implemented on Sina Weibo (Wikipedia 2015), existing accounts might contain personal information that can be linked to real identities of account holders. Therefore, we 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 we see in our experiments may differ in the context of well-established accounts.

Design implications & future work

Our 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 paper—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. We hope that our work in this paper will inspire designers and activists to come up with tools to promote freedom of speech and freedom of communication via social media under repressive regimes.

Conclusion

In this paper, we presented a non-deterministic algorithm to calculate homophones of Chinese words to transform censored keywords in social media to the ones that appear innocent in the eyes of censors. We conducted two experiments to see (1) how transformed social media posts perform compared to the original, unaltered posts and (2) whether native Chinese speakers have trouble understanding posts with homophone substitutions. The results were largely encouraging. Keyword transformations allow posts to be published on social media more than no transformation. We also found that the average age of transformed posts before they got removed was significantly higher than original posts that got removed. In our experiment with native Chinese speakers, nearly all of our transformed posts were easily understood by our workers. Workers who were not able to identify transformed and original keywords were more likely to have a hard time understanding the content of the posts. We also estimated that it would cost the social media operator 15 human-hours per day, per keyword to review false positive posts that match our homophones. Our approach to circumvent censorship can likely be assembled with other tools to build a real-time system for Chinese social media users to circumvent censorship.

Acknowledgments

We would like to thank our research group, colleagues that commented on early versions of this work, as well as our AMT workers for contributing to our research efforts.

References

Al-Ani, B.; Mark, G.; Chung, J.; and Jones, J. 2012. The egyptian blogosphere: A counter-narrative of the revolution. In *Proc. CSCW '12*.

Bamman, D.; O'Connor, B.; and Smith, N. 2012. Censorship and deletion practices in chinese social media. *First Monday* 17(3).

Chen, L.; Zhang, C.; and Wilson, C. 2013. Tweeting under pressure: Analyzing trending topics and evolving word choice on sina weibo. In *Proc. COSN '13*.

Clayton, R.; Murdoch, S. J.; and Watson, R. N. 2006. Ignoring the great firewall of china. In *Privacy Enhancing Technologies*, 20–35. Springer.

Fu, K.-w.; Chan, C.-h.; and Chau, M. 2013. Assessing censorship on microblogs in china: Discriminatory keyword analysis and the real-name registration policy. *IEEE Internet Computing* 17(3):42–50.

Huang, H.; Wen, Z.; Yu, D.; Ji, H.; Sun, Y.; Han, J.; and Li, H. 2013. Resolving entity morphs in censored data. In *Proc. ACL '13*.

King, G.; Pan, J.; and Roberts, M. E. 2013. How censorship in china allows government criticism but silences collective expression. *American Political Science Review* 107(02):326–343.

King, G.; Pan, J.; and Roberts, M. E. 2014. Reverse-engineering censorship in china: Randomized experimentation and participant observation. *Science* 345(6199):1251722.

Li, P., and Yip, M. C. 1996. Lexical ambiguity and context effects in spoken word recognition: Evidence from chinese. In *Proceedings of the 18th Annual Conference of the Cognitive Science Society*, 228–232.

MacKinnon, R. 2009. China's censorship 2.0: How companies censor bloggers. *First Monday* 14(2).

Wikipedia. 2015. Microblogging in China. Page Version ID: 637181125.

Wulf, V.; Misaki, K.; Atam, M.; Randall, D.; and Rohde, M. 2013. 'on the ground' in sidi bouzid: Investigating social media use during the tunisian revolution. In *Proc. CSCW '13*.

Zhang, B.; Huang, H.; Pan, X.; Ji, H.; Knight, K.; Wen, Z.; Sun, Y.; Han, J.; and Yener, B. 2014. Be appropriate and funny: Automatic entity morph encoding. In *Proc. ACL2014*.

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

Zuckerman, E. 2008. *My Heart's in Accra, March* 8.