Censorship and Deletion Practices in Chinese Social Media[*]

David Bamman Brendan O'Connor Noah A. Smith
School of Computer Science
Carnegie Mellon University
{dbamman,brenocon,nasmith}@cs.cmu.edu

Abstract

With Twitter and Facebook blocked in China, the stream of information from Chinese domestic social media provides a case study of social media behavior under the influence of active censorship. While much work has looked at efforts to prevent *access* to information in China (including IP blocking of foreign websites or search engine filtering), we present here the first large-scale analysis of political *content* censorship in social media, i.e., the active deletion of messages published by individuals.

In a statistical analysis of 56 million messages (212,583 of which have been deleted out of 1.3 million checked, more than 16%) from the domestic Chinese microblog site Sina Weibo, and 11 million Chinese-language messages from Twitter, we uncover a set a politically sensitive terms whose presence in a message leads to anomalously higher rates of deletion. We also note that the rate of message deletion is not uniform throughout the country, with messages originating in the outlying provinces of Tibet and Qinghai exhibiting much higher deletion rates than those from eastern areas like Beijing.

1 Introduction

Much research on Internet censorship has focused on only one of its aspects: IP and DNS filtering within censored countries of websites beyond their jurisdiction, such as the so-called "Great Firewall of China" (GFW) that prevents Chinese residents from accessing foreign websites such as Google and Facebook (FLOSS, 2011; OpenNet Initiative, 2009; Roberts et al., 2009), or Egypt's temporary blocking of social media websites such as Twitter during its protests in January 2011.

Censorship of this sort is by definition designed to be complete, in that it aims to prevent *all* access to such resources. In contrast, a more relaxed "soft" censorship allows access, but polices content. Facebook, for example, removes content that is "hateful, threatening, or pornographic; incites violence; or contains nudity or graphic or gratuitous violence" (Facebook, 2011). Aside from their own internal policies, social media organizations are also governed by the laws of the country in which they operate. In the United States, these include censoring the display of child pornography, libel, and media that infringe on copyright or other intellectual property rights; in China this extends to forms of political expression as well.

The rise of domestic Chinese microblogging sites has provided an opportunity to look at the practice of soft censorship in online social media in detail. Twitter and Facebook were blocked in China in July 2009 after riots in the western province of Xinjiang (Blanchard, 2009). In their absence, a number of domestic services have arisen to take their place; the largest of these is Sina Weibo,[1] with over 200 million users (Fletcher, 2011).

We focus here on leveraging a variety of information sources to discover and then characterize censorship and deletion practices in Chinese social media. In particular, we exploit three orthogonal sources of information: message deletion patterns on Sina Weibo; differential popularity of terms on Twitter vs. Sina; and terms that are blocked on Sina's search interface. Taken together, these information sources lead to three conclusions.

[*] Published in *First Monday* 17.3 (March 2012).

[1] http://www.weibo.com

1. External social media sources like Twitter (i.e., Chinese language speakers outside of China) can be exploited to detect sensitive phrases in Chinese domestic sites since they provide an uncensored stream for contrast, revealing what is *not* being discussed in Chinese social media.
2. While users may be prohibited from searching for specific terms at a given time (e.g., "Egypt" during the Arab Spring), content censorship allows users to publish politically sensitive messages, which are occasionally, though not always, deleted retroactively.
3. The rate of posts that are deleted in Chinese social media is not uniform across the entire country; provinces in the far west and north, such as Tibet and Qinghai, have much higher rates of deletion (53%) than eastern provinces and cities (ca. 12%).

Note that we are not looking at censorship as an abstraction (e.g., detecting keywords that are blocked by the GFW, regardless of the whether or not anyone uses them). By comparing social media messages on Twitter with those on domestic Chinese social media sites and assessing statistically anomalous deletion rates, we are identifying keywords that are currently highly salient in real public discourse. By examining the deletion rates of specific messages by real people, we can see censorship in action.

2 Internet Censorship in China

MacKinnon (2011) and the OpenNet Initiative (2009) provide a thorough overview of the state of Internet filtering in China, along with current tactics in use to sway public discourse online, including cyberattacks, stricter rules for domain name registration, localized disconnection (e.g., Xinjiang in July 2009), surveillance, and astroturfing (MacKinnon, 2011; OpenNet Initiative, 2009; Bandurski, 2008).

Prior technical work in this area has largely focused on four dimensions. In the security community, a number of studies have investigated **network filtering** due to the GFW, discovering a list of blacklisted keywords that cause a GFW router to sever the connection between the user and the website they are trying to access (Crandall et al., 2007; Xu et al., 2011; Espinoza and Crandall, 2011); in this domain, the Herdict project[2] and Sfakianakis et al. (2011) leverage a global network of users to report unreachable URLs. Villeneuve (2008b) examines the **search filtering** practices of Google, Yahoo, Microsoft and Baidu in China, noting extreme variation between search engines in the content they censor, echoing earlier results by the Human Rights Watch (2006). Knockel et al. (2011) and Villeneuve (2008a) reverse engineer the TOM-Skype chat client to detect a list of sensitive terms that, if used, lead to **chat censorship**. MacKinnon (2009) evaluates the **blog censorship** practices of several providers, noting a similarly dramatic level of variation in suppressed content, with the most common forms of censorship being keyword filtering (not allowing some articles to be posted due to sensitive keywords) and deletion after posting.

This prior work strongly suggests that domestic censorship in China is deeply fragmented and decentralized. It uses a porous network of Internet routers usually (but not always) filtering the worst of blacklisted keywords, but the censorship regime relies more heavily on domestic companies to police their own content under penalty of fines, shutdown and criminal liability (Crandall et al., 2007; MacKinnon, 2009; OpenNet Initiative, 2009).

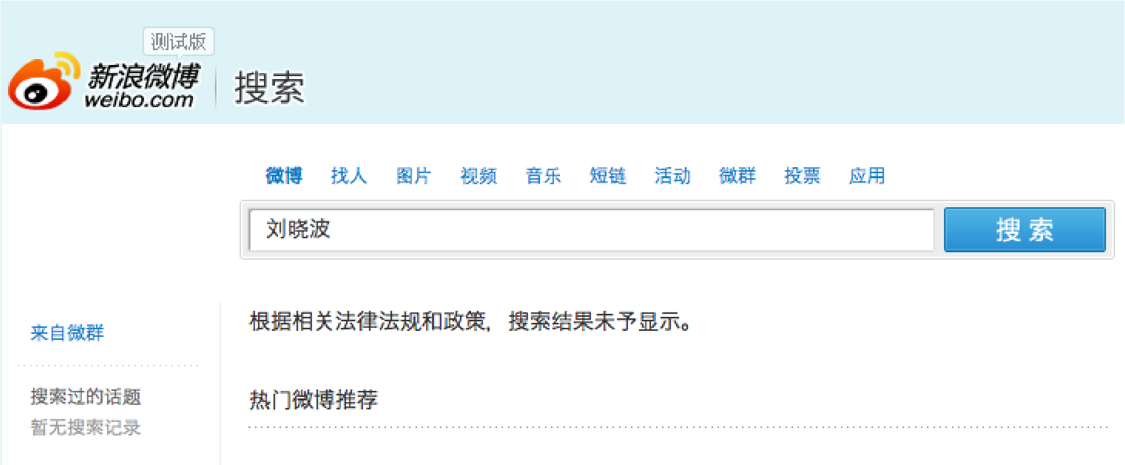
3 Microblogs

Chinese microblogs have, over the past two years, taken front stage in this debate, both in their capacity to virally spread information and organize individuals, and in several high-profile cases of government control. One of the most famous of these occurred in October 2010, when a 22-year-old named Li Qiming killed one and injured another in a drunk driving accident at Hebei University. His response after the accident—"Go ahead, sue me if you dare. My dad is Li Gang!" (deputy police chief in a nearby district)—rapidly spread on social media, fanning public outrage at government corruption and leading censors to instruct media sources to stop all "hype regarding the disturbance over traffic at Hebei University" (Qiang, 2011; Wines, 2010). In December 2010, Nick Kristof of the *New York Times* opened an account on Sina Weibo to test its level of censorship (his first posts were "Can we talk about Falun Gong?" and "Delete my weibos if you dare! My dad is Li Gang!" (Kristof, 2011b). A post on Tiananmen Square was deleted by moderators within twenty minutes; after attracting the wider attention of the media, his entire user account was shut down as well (Kristof, 2011a).

Beyond such individual stories of content censorship, there are a far greater number of reports of *search censorship*, in which users are prohibited from searching for messages containing certain keywords. An example of this is shown

[2] `http://www.herdict.org`

in Figure 1, where an attempt to search for "Liu Xiaobo" on October 30, 2011 is met with a message stating that, "according to relevant laws, regulations and policies, the search results were not shown." Reports of other search terms being blocked on Sina Weibo include "Jasmine" (sc. Revolution) (Epstein, 2011) and "Egypt" (Wong and Barboza, 2011) in early 2011, "Ai Weiwei" on his release from prison in June 2011 (Gottlieb, 2011), "Zengcheng" during migrant protests in that city in June 2011 (Kan, 2011), "Jon Huntsman" after his attendance at a Beijing protest in February 2011 (Jenne, 2011), "Chen Guangcheng" (jailed political activist) in October 2011 (Spegele, 2011) and "Occupy Beijing" and several other place names in October 2011 following the "Occupy Wall Street" movement in the United States (Hernandez, 2011).



刘晓波

根据相关法律法规和政策，搜索结果未予显示。

Figure 1: Results of attempted search for *Liu Xiaobo* (political dissident and Nobel prize winner) on Sina Weibo: "According to relevant laws, regulations and policies, the search results were not shown."

4 Message Deletion

Reports of message deletion on Sina Weibo come both from individuals commenting on their own messages (and accounts) disappearing (Kristof, 2011a), and from allegedly leaked memos from the Chinese government instructing media to remove all content relating to some specific keyword or event (e.g., the Wenzhou train crash) (CDT, 2011). Charles Chao, the CEO of Sina Weibo, reports that the company employs at least one hundred censors, though that figure is thought to be a low estimate (Epstein, 2011). Manual intervention can be seen not only in the deletion of sensitive messages containing text alone, but also in those containing subversive images and videos as well (Larmer, 2011).

To begin exploring this phenomenon, we collected data from Sina Weibo over the period from June 27 to September 30, 2011. Like Twitter and other social media services, Sina provides developers with open APIs on which to build services, including access methods to timeline and social graph information. In order to build a dataset, we queried the public timeline at fixed intervals to retrieve a sample of messages. Over the three month period, this led to a total collection of 56,951,585 messages (approximately 600,000 messages per day).

Each message in our collection was initially written and published at some point between June 27 and September 30, 2011. For each of these messages, we can check, using the same API provided to developers, whether the message exists and can be read today, or if it has been deleted at some point between now and its original date of publication. If it has been deleted, Sina returns the message "`target weibo does not exist`."

In late June/early July 2011, rumors began circulating in the Chinese media that Jiang Zemin, general secretary of the Communist Party of China from 1989 to 2002, had died. These rumors reached their height on July 6, with reports in *The Wall Street Journal*, *The Guardian* and other western media sources that Jiang's name (江泽民) had been blocked in searches on Sina Weibo (Chin, 2011; Branigan, 2011).

If we look at all 532 messages published during this time period that contain the name *Jiang Zemin* (Figure 2), we note a striking pattern of deletion: on July 6, the height of the rumor, 64 of the 83 messages containing that name were deleted (77.1%); on July 7, 29 of 31 (93.5%) were deleted.

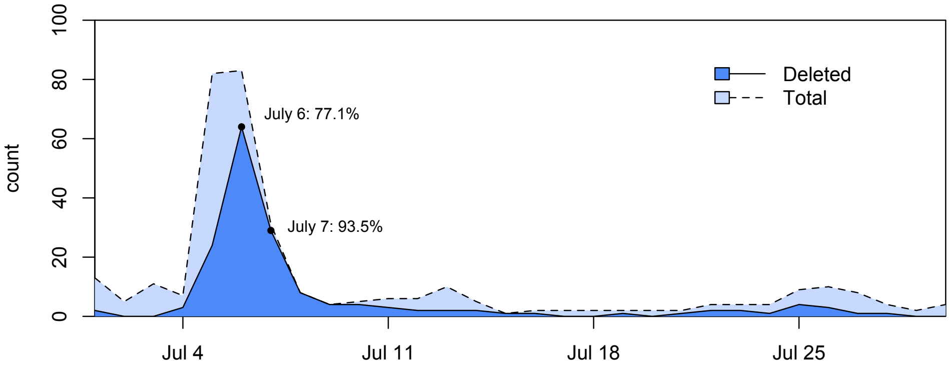


Figure 2: Number of deleted messages and total messages containing the phrase *Jiang Zemin* on Sina Weibo.

Messages can of course be deleted for a range of reasons, and by different actors: social media sites, Twitter included, routinely delete messages when policing spam; and users themselves delete their own messages and accounts for their own personal reasons. But given the abnormal pattern exhibited by *Jiang Zemin* we hypothesize that there exists a set of terms that, given their political polarity, will lead to a relatively higher rate of deletion for all messages that contain them.

4.1 Term Deletion Rates

In this section, we develop our first sensitive term detection procedure: collect a uniform sample of messages and whether they are deleted, then rank terms by deletion rate, while controlling for statistical significance with the method of *false discovery rate* (Benjamini and Hochberg, 1995).

We first build a deleted message set by checking whether or not messages originally published between June 30 and July 25, 2011 still existed three months later (i.e., messages published on June 30 were checked for existence on October 1; those published on July 25 were checked on October 26). We wish to remove spam, since spam is a major reason for message deletion, but we are interested in politically-driven message deletions. We filtered the entire dataset on three criteria: (1) duplicate messages that contained exactly the same Chinese content (i.e., excluding whitespace and alphanumerics) were removed, retaining only the original message; (2) all messages from individuals with fewer than five friends and followers were removed; (3) all messages with a hyperlink (`http`) or addressing a user (`@`) were removed if the author had fewer than one hundred friends and followers. Over all the data published between June 30 and July 25, we checked the deletion rates for a random sample of 1,308,430 messages, of which 212,583 had been deleted, yielding a baseline message deletion rate δ_b of 16.25%.

Next, we extracted terms from the messages. In Chinese, the basic natural language processing task of identifying words in text can be challenging due to the absence of whitespace separating words (Sproat and Emerson, 2003). Rather than attempting to make use of out-of-domain word segmenters that may not generalize well to social media, we first constructed a Chinese-English dictionary as the union of the open source CC-CEDICT dictionary[3] and all entries in the Chinese-language Wikipedia[4] that are aligned to pages in English Wikipedia; we use the English titles to automatically derive Chinese-English translations for the terms. Using Wikipedia substantially increases the number of named entities represented. The full lexicon has 255,126 unique Chinese terms. After first transforming any traditional characters into their simplified equivalents, we then identified words in a message as all character n-grams up to length 5 that existed in the lexicon (this includes overlaps and overgenerates in some cases).

We then estimate a *term deletion rate* for every term w in the vocabulary,

$$\delta_w \equiv P(\text{message becomes deleted} \mid \text{message contains term } w) = \frac{d_w}{n_w} \qquad (1)$$

where d_w is the number of deleted messages containing w and n_w is the total number of messages containing w. It is misleading to simply look at the terms that have the highest deletion rates, since rarer terms have much more variable δ_w given their small sample sizes. Instead, we would like to focus on terms whose deletion rates are both high as well as *abnormally* high given the variability we expect due to sampling. We graphically depict these two factors in Figure 3.

[3] `http://www.mdbg.net/chindict/chindict.php?page=cedict`

[4] `http://zh.wikipedia.org`

Every point is one term; its overall message count is shown on the x-axis, versus its deletion rate δ_w on the y-axis. For every message count, we compute extreme quantiles of the binomial null hypothesis that messages are randomly deleted at the base rate of 16.25%. For example, for a term that occurs in 10 messages, in 99.9% of samples, 6 or fewer of them should be deleted under the null hypothesis; i.e., $P_{null}(D \leq 6 \mid N = 10) < 0.999 < P_{null}(D \leq 7 \mid N = 10)$, where P_{null} denotes the null hypothesis distribution, D is the number of deleted messages (a random variable), and N is the total number of messages containing the term (another random variable). Therefore in Figure 3, at $N = 10$ the upper line is plotted at 0.6.

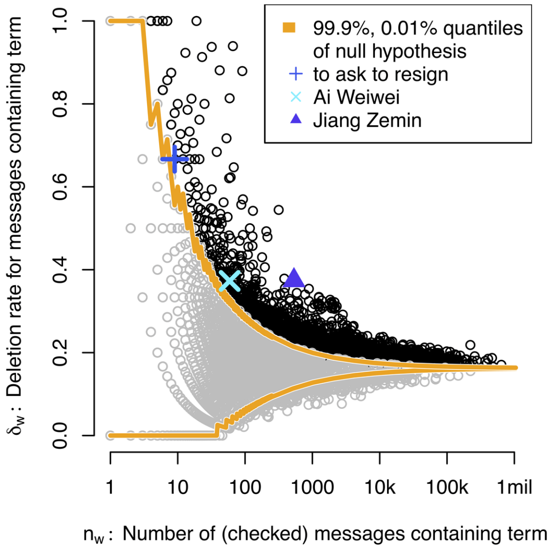


Figure 3: Deletion rates per term, plotting a term's overall frequency against the probability a message it appears in is deleted. One point per term. Black points have $p_w < 0.001$.

When terms are more frequent, their observed deletion rates should naturally be closer to the base rate. This is illustrated as the quantile lines coming together at higher frequencies.[5] As we might expect, the data also show that higher frequency terms have deletion rates closer to the base rate. However, terms' deletion rates vary considerably more than the null hypothesis, and substantially more in the positive high-deletion direction. If the null hypothesis were true, only one in 1,000 terms would have deletion rates above the top orange line. But 4% of our terms have deletion rates in this range, indicating that deletions are substantially non-random conditional on textual content.

That fact alone is unremarkable, but this analysis gives a useful way to filter the set of terms to interesting ones whose deletion rates are abnormally high. For every term, we calculate its deletion rate's one-tailed binomial p-value,

$$\begin{aligned} p_w &\equiv P_{null}(D \geq d_w \mid N = n_w) \\ &= 1 - \text{BinomCDF}(d_w; n_w, \delta = 0.1625) \end{aligned}$$

and use terms with small p_w as promising candidates for manual analysis. How reliably non-null are these terms? We are conducting tens of thousands of simultaneous hypothesis tests, so must apply a multiple hypothesis testing correction. We calculate the *false discovery rate* $P(null \mid p_w < p)$, the expected proportion of false positives within the set of terms passing a threshold p. Analyzing the $p_w < 0.001$ cutoff, the Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995) gives an upper bound on FDR of

$$\text{FDR}_{p_w < .001} < \frac{P_{null}(p_w < p)}{\hat{P}(p_w < p)} = \frac{0.001}{0.040} = 2.5\% \tag{2}$$

where $\hat{P}$ is the empirically estimated distribution of deletion rate p-values; i.e., how many points are beyond the orange line. The ratio simply reflects how much more often extreme values happen, in contrast to chance. Since we expect fewer than 1 out of 40 of these terms could have been generated at random, they are reasonable candidates for further analysis. We could also use more stringent thresholds if desired:

[5] For this reason, this statistical visualization is known as a *funnel plot* (Spiegelhalter, 2005).

threshold	FDR	# selected terms (out of 75,917 total)
$p_w < 10^{-3}$	0.025	3,046
$p_w < 10^{-4}$	0.003	2,181
$p_w < 10^{-5}$	0.0004	1,715

False discovery rate control is widespread in bioinformatics, since it gracefully handles tens of thousands of simultaneous hypothesis tests (unlike the Bonferroni correction). It is analogous to ($1-$precision) and quite different than the Type I or II error rates in classical hypothesis testing—and arguably more meaningful for large-scale inference (Efron, 2010; Storey and Tibshirani, 2003).

Finally, we also performed additional deletion checks on another set of 33,363 messages that contained one of the 295 sensitive terms described in §7. We used this data to calculate deletion rates for these terms (incorporated in Figure 3), giving better statistical confidence than using their rates in the uniform sample, since there was substantially more data per term. This targeted sampling is more biased, of course, but is useful since API limits restrict the total number of deletion checks we can perform.

4.2 Analysis of Highly Deleted Terms

We qualitatively analyzed the most highly deleted terms that passed the $p_w < 0.001$ threshold. These terms span a range of topics: several of the most frequently deleted terms appear in messages that are clearly spam (including movie titles and actors), and it is necessary to discard one-character words (while these are valid words in our dictionary, they are often partial words when analyzed in context). All terms mentioned in this section have deletion rates in the 50%–100% range.

Several interesting categories emerge. One is the clear presence of known politically sensitive terms, such as 方滨兴 (Fang Binxing, the architect of the GFW), 真理部 ("Ministry of Truth," a reference to state propaganda), and 法轮功 (Falun Gong, a banned spiritual group). Another is a set of terms that appear to have become sensitive due to changing real-world events. One example of this is the term 请辞 (to ask someone to resign); deleted messages containing this term call for the resignation of Sheng Guangzu, the Minister of Railways, following the Wenzhou train crash in July (Chan and Duncan, 2011). Another example is the term 两会 (two meetings): this term primarily denotes the joint annual meeting of the National People's Congress and the Chinese People's Political Consultative Conference, but also emerged as a code word for "planned protest" during the suppression of pro-democracy rallies in February and March 2011 (Kent, 2011).

The most topically cohesive of these are a set of terms found in messages relating to the false rumor of iodized salt preventing radiation poisoning following the Fukushima nuclear disaster of March 2011 (Burkitt, 2011). Highly deleted terms in this category include 防核 (nuclear defense/protection), 碘盐 (iodized salt), and 放射性碘 (radioactive iodine). Unlike other terms whose political sensitivity is relatively static, these are otherwise innocuous terms that become sensitive due to a dynamic real-world event: instructions by the Chinese government not to believe or spread the salt rumors (Xin et al., 2011). Given recent government instructions to social media to quash false rumors in general (Chao, 2011), we believe that these abnormally high deletions constitute the first direct evidence for suppression of this rumor as well. In addition to specific messages relating to salt, we also observe more general news and commentary about the nuclear crisis appearing more frequently in deleted messages, leading to abnormally high deletion rates for terms such as "nuclear power plant," "nuclear radiation," and "Fukushima."

In the absence of external corroborating evidence (such as reports of the Chinese government actively suppressing salt rumors, as above), these results can only be suggestive, since we can never be certain that a deletion is due to the act of a censor rather than other reasons. In addition to terms commonly associated with spam, some terms appear frequently in cases of clear personal deletions; examples include the names of several holidays (e.g. 元宵节, the Lantern Festival), and expressions of condolences (节哀顺变). Given this range of deletion reasons, we turn to incorporating other lexical signals to focus on politically sensitive keywords.

5 Twitter vs. Sina Comparison

The second source of information that we can use to filter the highly-deleted term list is a comparison of word frequency on Twitter vs. Sina Weibo. Since Twitter is not reported to censor the stream of data from its users globally,[6] it may provide a baseline against which to measure global attention to a certain topic. We build a dataset using Twitter's streaming API; since a small fraction of Twitter's public timeline is comprised of Chinese-language tweets, we

[6] Twitter does retroactively filter certain messages in response to specific local demands (Twitter, 2012).

identified the 10,000 most frequent users in the gardenhose sample over the period June 1–24, 2011 writing tweets in Chinese not containing `http` or `www` (to filter spammers). These users are a mix of Western news sources (with tweets in Chinese), overseas Chinese speakers, and users within China accessing Twitter via proxy networks; as such, they may reflect a more Western-oriented bias than the uniform sample of mainland Chinese users who use Sina Weibo. We then retrieved the public streams of these 10,000 users via Twitter's streaming API. Over the three month period of data collection, this resulted in a data set of 11,079,704 tweets.

Jiang Zemin again provides a focal point: a trend analysis reveals a dramatic increase in the frequency of mention of Jiang's name on Twitter and a much smaller increase on Sina. At the height on July 6, Jiang's name appeared on Twitter with a relative document frequency of 0.013, or once every 75 messages, two orders of magnitude more frequently than Sina (once every 5,666 messages). Twitter is clearly on the leading edge of these rumors, with reports about Jiang's declining health appearing most recently on June 27 and the first rumors of his death appearing on June 29. We note the same pattern emerging with other terms that have historically been reported to be sensitive, including 艾未未 (Ai Weiwei) and 刘晓波 (Liu Xiaobo), as shown in Figure 4.

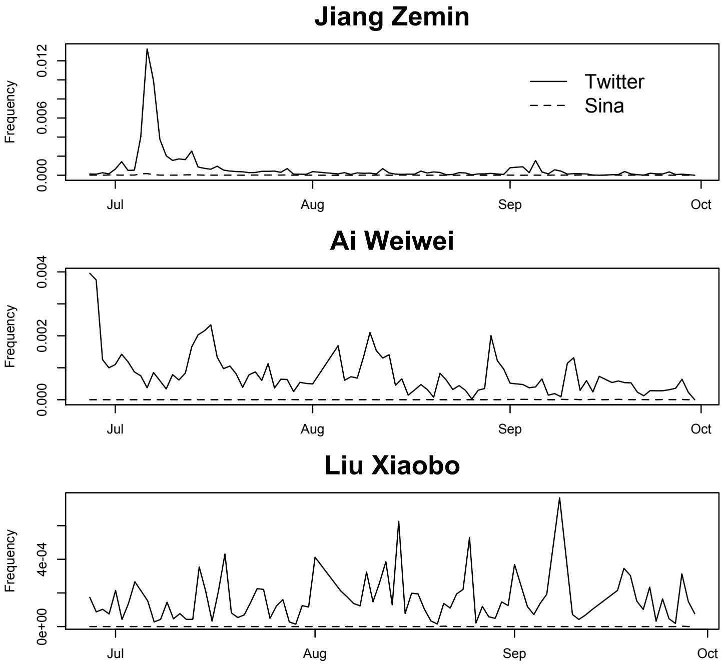


Figure 4: Time series for *Jiang Zemin*, *Ai Weiwei* and *Liu Xiaobo* on Twitter (solid) and Sina (dashed), showing message frequency of each term by day. All terms appear several orders of magnitude more frequently on Twitter (where they show typical variation over time) than Sina.

This suggests a hypothesis: whatever the source of the difference, the objective discrepancy in term frequencies between Twitter and the domestic social media sites may be a productive source of information for automatically identifying which terms are politically sensitive in contemporary discourse online.

To test this, we rank every term in our vocabulary by its comparative log-likelihood ratio between sources: the frequency of the term on Twitter over its frequency on Sina.

$$\text{LLR}_w = \log \frac{P(w \mid source = \text{Twitter})}{P(w \mid source = \text{Sina})} \tag{3}$$

6 Search Blocking

To test the viability of this approach for locating sensitive keywords, we ranked all terms by their log likelihood scores and checked whether each of the top 2,000 terms was blocked by the search interface on Sina Weibo (as in Figure 1). While this evaluation can only confirm terms that are governed by hard censorship (not the soft censorship we are interested in), it does provide confirmation that such terms are indeed sensitive.

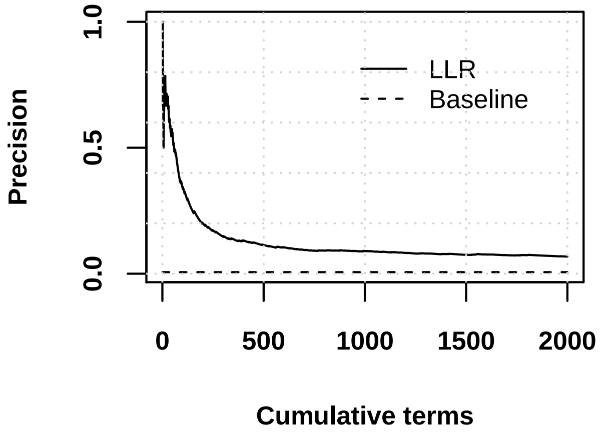


Figure 5: Search block precision by the number of ranked terms tested.

†	term	gloss
†	何德普	He Depu
†	刘晓波	Liu Xiaobo
	北京市监狱	Beijing Municipal Prison
†	零八宪章	Charter 08
	廖廷娟	Liao Tingjuan
	廖筱君	Liao Hsiao-chun
†	共匪	communist bandit
†	李洪志	Li Hongzhi, founder of the Falun Gong spiritual movement
†	柴玲	Chai Ling
†	方滨兴	Fang Binxing
†	法轮功	Falun Gong
†	大纪元	*Epoch Times*
†	刘贤斌	Liu Xianbin
†	艾未未	Ai Weiwei, Chinese artist and activist
	王炳章	Wang Bingzhang
	非公式	unofficial/informal (Japanese)
†	魏京生	Wei Jingsheng, Beijing-based Chinese dissident
	唐柏桥	Tang Baiqiao
†	鲍彤	Bao Tong
†	退党	to withdraw from a political party

Table 1: Search block status on October 24, 2011 of the 20 terms with the highest Twitter/Sina log likelihood ratio scores. Search blocked terms are noted with a †.

Figure 5 displays the precision (the number of terms found to be blocked on search divided by the number of terms checked) for the top x terms with the highest log likelihood scores. The results show a heavy tail, with 70.0% of the top 20 terms with the highest log-likelihood ratios being blocked, 56.0% of the top 50 terms, 34.0% of the top 100, 11.4% of the top 500, 9.0% of the top 1,000, and 6.8% of the top 2,000 (yielding a total of 136 search-blocked terms). To establish a baseline in comparison, we sampled 1,000 terms from n-grams of length one to five (200 terms uniformly at random for each n-gram length) after discarding the most frequent 5% and least frequent 5% within each length. The baseline rate of a randomly drawn term being deleted is 0.6% (1.5% for unigrams, 0% for bigrams, 0% for trigrams, 1.5% for 4-grams, 0% for 5-grams).

Table 1 lists the top twenty terms with the highest LLR score, along with their search block status on October 24, 2011.[7] The terms that are much more heavily discussed on Twitter than on Sina and also blocked on Sina's search

[7] On March 1, 2012, two of these terms (Fang Binxing and Ai Wei Wei) were no longer blocked, revealing the dynamic nature of the censorship system.

interface include political dissidents such as He Depu, Liu Xiaobo, and Ai Weiwei, terms and people associated with the Falun Gong movement, and western news media (*Epoch Times*). Terms more frequently discussed on Twitter than Sina and not blocked include innocuous terms such as Taiwanese television personalities and 非公式 ("unofficial," a predominantly Japanese word that appears in Japanese-language tweets), but also terms that are politically sensitive, including the Beijing Municipal Prison (where several political prisoners are held) and pro-democracy activist Wang Bingzhang.

These results corroborate earlier reports of individual search terms being blocked on Sina Weibo (Epstein, 2011; Wong and Barboza, 2011; Gottlieb, 2011; Kan, 2011; Jenne, 2011) and provide an avenue for automatically detecting the rise of new terms in the future. More importantly, by being blocked on Sina's search interface, these terms are confirmed to be politically sensitive, and can act as a filter for the set of terms with anomalously higher rates of deletion found in section 4.

7 Deletion Rates of Politically Sensitive Terms

Section 4 described our efforts at looking at term deletion rates in a uniform sample of all messages. With a set of known politically sensitive terms discovered through the process above, we can now filter those results and characterize the deletion of messages on Sina Weibo that are due not to spam, but to the presence of known politically sensitive terms within them.

The 136 terms from the Twitter/Sina comparative LLR list that are blocked on Sina's search interface are inherently politically sensitive by virtue of being blocked. To this list we also add two sets of terms from previous work that have been shown to be politically sensitive as well: (1) a list of blacklisted keywords discovered via networking filtering by Crandall et al. (2007);[8] and (2) a list of blacklisted terms manually compiled on Wikipedia (Wikipedia, 2011). This results in a total of 295 politically sensitive terms.

We identified every message containing each term in our full dataset of 56 million messages, and checked whether that message had been deleted. 33,363 messages were found to contain at least one of those sensitive terms, and 5,811 of those messages (17.4%) had been deleted.

Table 2 lists the results of this analysis. 17 terms known to be politically sensitive are deleted at rates significantly higher than the baseline, chosen such that the upper bound on the false discovery rate is 2.5% (we expect less than 1 in 40 to have resulted from chance).

δ_w	deletions	total	term	gloss	source(s)
1.000	5	5	方滨兴	Fang Binxing	T
1.000	5	5	真理部	Ministry of Truth	T
0.875	7	8	法轮功	Falun Gong	T
0.833	5	6	共匪	communist bandit	T,W
0.717	38	53	盛雪	Sheng Xue	C
0.500	13	26	法轮	Falun	T,C,W
0.500	16	32	新语丝	*New Threads*	C
0.379	145	383	反社会	antisociety	C
0.374	199	532	江泽民	Jiang Zemin	T,C,W
0.373	22	59	艾未未	Ai Weiwei	T
0.273	41	150	不为人知的故事	"The Unknown Story"	W
0.257	119	463	流亡	to be exiled	W
0.255	82	321	驾崩	death of a king or emperor	T
0.239	120	503	浏览	to browse	C
0.227	112	493	花花公子	Playboy	C,W
0.226	167	740	封锁	to blockade	W
0.223	142	637	大法	(sc. Falun) Dafa	W

Table 2: Sensitive terms with statistically significant higher rates of message deletion ($p < 0.001$). Source designates whether the sensitive term originates in our Twitter LLR list (T), Crandall et al. (2007) (C), or Wikipedia (Wikipedia, 2011) (W).

[8] The exact list used can be found at `http://www.conceptdoppler.org/GETRequestBlocked18June.html`.

Among this set, the most heavily deleted include Fang Binxing (the architect of the Great Firewall) (Chao, 2010), Falun Gong, political activists Sheng Xue and Ai Weiwei, foreign news media (*New Threads*), Jiang Zemin, and terms related to pornography (Playboy, "pornographic"). One observation that emerges from this analysis is that while messages containing these 17 terms are deleted at statistically higher rates than messages which do not, the practice of social media censorship is far more nuanced than a simple blacklist might suggest. While most of these terms are officially blocked on Sina's search interface, very few of them are deleted with 100% coverage (indeed, only terms that occur a handful of times are deleted completely). If we look at a list of terms that have been previously shown to be blacklisted (with respect to the GFC), we see that many of those terms are freely used in messages on Sina Weibo, and in fact still can be seen there at this writing. Table 3 lists a sample of terms from Crandall et al. (2007) that appear in over 100 messages in our sample and are *not* deleted at statistically higher rates. Many of these terms cannot be searched for via Sina's interface, but frequently appear in actual messages.

†	δ_w	deletions	total	term	gloss
†	0.20	88	443	中宣部	Central Propaganda Section
†	0.20	24	120	藏独	Tibetan independence (movement)
	0.19	30	154	民联	Democratic Alliance
†	0.18	132	733	迫害	to persecute
	0.18	124	686	酷刑	cruelty/torture
	0.18	80	457	钓鱼岛	Senkaku Islands
†	0.18	28	153	太子党	Crown Prince Party
†	0.17	102	592	法会	Falun Gong religious assembly
†	0.17	88	526	纪元	last two characters of Epoch Times
	0.17	56	333	民进党	DPP (Democratic Progressive Party, Taiwan)
	0.16	142	863	洗脑	brainwash
†	0.16	42	256	我的奋斗	Mein Kampf
†	0.15	83	567	学联	Student Federation
	0.15	32	208	高瞻	Gao Zhan
	0.14	51	360	无界	first two characters of circumventing browser
	0.14	36	250	正念	correct mindfulness
†	0.14	28	198	天葬	sky burial
	0.14	17	122	文字狱	censorship jail
	0.13	90	677	经文	scripture
†	0.12	91	732	八九	89 (the year of the Tiananmen Square Protest)
†	0.12	67	564	看中国	watching China, an Internet news website
†	0.11	35	310	明慧	Ming Hui (website of Falun Gong)
†	0.10	56	582	民运	democracy movement

Table 3: Deletion rates of terms from Crandall et al. (2007), previously reported to be blocked by the GFC, that appear frequently (over 100 times) in our sample. Terms that are currently blocked on Sina's search interface are noted with a †.

To determine if there is some principled reason behind the discrepancy in deletion rates, we further looked at two properties of the message sets that might, *a priori*, be significant factors in determining whether they were deleted: the potential impact of a message, measured by the number of times it was rebroadcast (or "retweeted" on Twitter) and the number of followers of their authors; and the message content itself (i.e., whether the message is expressing a positive or negative sentiment toward the sensitive topic).

Message Impact Prior work has shown that rebroadcasting accounts for a large part of user activity on Sina Weibo, especially in its function for developing trends (Yu et al., 2011). We might suspect that if a politically sensitive message is being heavily rebroadcast, it may be more likely to be deleted. While our dataset consists of original messages only, the Sina API also provides information on the number of times any given message was rebroadcast and commented upon (even deleted messages). We gathered this information for all 33,363 messages that contain at least one sensitive keyword: while most messages had never been rebroadcast (leading to a median of 0 for both deleted and not-deleted messages), 14.7% of deleted messages had been rebroadcast at least once, along with 23.2% of messages that had not been deleted. When excluding outliers from both sets that were rebroadcast over 100 times, the difference between mean rebroadcast counts for both sets (0.9368 for the deleted set, 0.9518 for the not deleted set) is not statistically significant.

Similarly, we might suspect that politically sensitive messages from authors with more followers are more likely to be deleted than those with fewer followers. An analysis of the 33,363 messages again shows no support for this: the mean number of followers for authors with deleted messages is 270.9 (median = 138) compared with 287.8 (median = 132) for authors of messages that were not deleted. In this case, we cannot reject the null hypothesis that messages with sensitive terms are equally likely to be deleted regardless of the number of followers or rebroadcast count.

Message Content One view of censorship is that any message with a politically sensitive phrase, either *pro* or *con*, is itself inherently politically sensitive only by virtue of containing that phrase. This is the implicit assumption behind prohibiting users from searching for certain terms – users on Sina Weibo currently cannot search for any messages mentioning Liu Xiaobo or Ai Weiwei, even if those messages are negatively oriented toward them. One possible explanation then for why *all* messages containing such sensitive phrases are not deleted is that only those expressing support for the politically sensitive term are deleted; those with views aligned with those of the censors are permitted to stay. To evaluate this hypothesis, we manually analyzed a small dataset of all 59 messages containing the phrase "Ai Weiwei," classifying each as *positive* (i.e., expressing sentiment supporting him), *negative* (expressing negative attitude toward him), *neutral* (stating a fact, such as the location of an art exhibit), and *unknown* (for ambiguous cases). With such a small sample, we can only offer an existence proof: of the 16 unambiguously positive messages toward Ai Weiwei (expressing support for him or criticism of the Chinese government with respect to its treatment of him), only 5 were deleted; 11 remain at the time of this writing.

The existence of such messages may suggest a random component to deletion (e.g., due to sampling); but here again, we cannot establish an explanation for why some messages containing politically sensitive terms are deleted and others are not. One area where we can see a sharp distinction between the two datasets, however, is where they originate geographically.

8 Geographic Distribution

As with Twitter, messages on Sina Weibo are attended with a range of metadata features, including free-text categories for user name and location and fixed-vocabulary categories for gender, country, province, and city. While users are free to enter any information they like here, true or false, this information can in the aggregate enable us to observe large-scale geographic trends both in the overall message pattern (Eisenstein et al., 2010, 2011; O'Connor et al., 2010; Wing and Baldridge, 2011) and in rates of deletion.

To conduct this analysis, we looked at all 1,308,430 messages whose deletion status we had checked, extracted their provinces from the metadata and estimated the probability of a message being deleted given the province that it came from as simply the count of messages from a province that are deleted divided by the total number of messages originating from that province.

$$\delta_p \equiv P(\text{message becomes deleted} \mid \text{message originates in province } p) \tag{4}$$

Figure 6 and Table 4 present the results of this geographic analysis. Messages that self-identify as originating from the outlying provinces of Tibet, Qinghai, and Ningxia are deleted at phenomenal rates: up to 53% of all messages originating from Tibet are deleted, compared with 12% from Beijing and 11.4% for Shanghai. We might suspect that higher rates of deletion in these areas may be connected with their histories of unrest (especially in Tibet, Qinghai, Gansu, and Xinjiang), but there are several possible alternative explanations: Sina censors may be deleting messages with greater frequency due to increased attention to these areas, perhaps enabled by the comparatively small volume of messages originating from them—the deletion rate by province is negatively rank correlated with message volume (Kendall's $\tau = -.73$); an alternative explanation is that users themselves are self-censoring at higher rates (Shklovski and Kotamraju, 2011).

One hypothesis that we have discounted is that the increased deletions are due to spam, since we observe similar deletion patterns when looking only at the subset of 33,363 messages that contain at least one of the 295 known politically sensitive keywords. In this subset, the same three provinces have the highest deletion rates (Ningxia with an overall deletion rate of 57.8%, Tibet 50%, and Qinghai 47.7%), while the lowest is Beijing (12.2%). The overall per-province deletion rates are largely similar (Kendall's $\tau = 0.77$, Pearson $r = 0.94$).

To explore this discrepancy further, we analyze the most characteristic words in each province – those words that are used more frequently in one province than in others. For each province, we find words having the highest *pointwise mutual information*,

$$\arg\max_w \text{PMI}(w; \text{province}) = \log \frac{P(w \mid \text{province})}{P(w)} \tag{5}$$

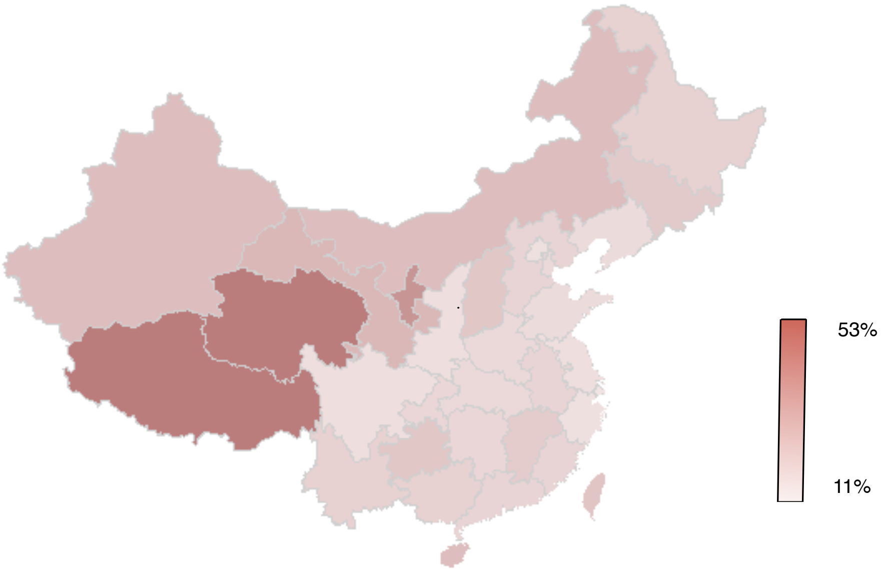


Figure 6: Deletion rates by province (darker = higher rates of deletion). This map visualizes the results shown in Table 4.

We restrict attention to words appearing in at least 50 messages in our 1.3 million message sample. For messages originating in Beijing, outside China, Qinghai, and Tibet, we present the top three terms overall, and the top politically sensitive terms in each region along with their PMI rank.

- Beijing: (1) 西直门 (Xizhimen neighborhood of Beijing); (2) 望京 (Wangjing neighborhood of Beijing); (3) 回京 (to return to the capital)
  ▷ (410) 钓鱼岛 (Senkaku/Diaoyu Islands)
- Outside China: (1) 多伦多 (Toronto); (2) 墨尔本 (Melbourne); (3) 鬼佬 (foreigner [Cantonese])
  ▷ (632) 封锁 (to blockade/to seal off); (698) 人权 (human rights)
- Qinghai: (1) 西宁 (Xining [capital of Qinghai]); (2) 专营 (special trade/monopoly); (3) 天谴 (divine retribution).
  ▷ (331) 独裁 (dictatorship); (803) 达赖喇嘛 (Dalai Lama)
- Tibet: (1) 拉萨 (Lhasa [capital of Tibet]); (2) 集中营 (concentration camp); (3) 贱格 (despicable)
  ▷ (50) 达赖喇嘛 (Dalai Lama); (108) 迫害 (to persecute)

Here the most characteristic terms in each province naturally tend to be locations within each area; while politically sensitive terms have weaker correlations with each region (e.g., the first known politically sensitive term in Beijing has only the 410th highest PMI), we do note the mention of the Dalai Lama in both Tibet and Qinghai, persecution in Tibet, and human rights as a general concern primarily outside China.

9 Conclusion

Chinese microblogging sites like Sina Weibo, Tencent, Sohu and others have the potential to change the face of censorship in China by requiring censors to police the content of over 200 million producers of information. In this large-scale analysis of deletion practices in Chinese social media, we showed that what has been suggested anecdotally by individual reports is also true on a large scale: there exists a certain set of terms whose presence in a message leads to a higher likelihood for that message's deletion. While a direct analysis of term deletion rates over all messages reveals a mix of spam, politically sensitive terms, and terms whose sensitivity is shaped by current events, a comparative analysis of term frequencies on Twitter vs. Sina provides a method for identifying suppressed political terms that are currently salient in global public discourse. By revealing the variation that occurs in censorship both in response to current events and in different geographical areas, this work has the potential to actively monitor the state of social media censorship in China as it dynamically changes over time.

	$\delta_{uniform}$	total$_{uniform}$	$\delta_{sensitive}$	total$_{sensitive}$
Tibet	0.530 ±0.01998	2406	0.500 ±0.106	86
Qinghai	0.521 ±0.01944	2542	0.477 ±0.104	88
Ningxia	0.422 ±0.01826	2880	0.578 ±0.097	102
Macau	0.321 ±0.01817	2910	0.400 ±0.101	95
Gansu	0.285 ±0.01365	5156	0.301 ±0.074	176
Xinjiang	0.270 ±0.01203	6638	0.304 ±0.070	194
Hainan	0.265 ±0.00932	11068	0.316 ±0.071	193
Inner Mongolia	0.263 ±0.01232	6332	0.278 ±0.068	209
Taiwan	0.239 ±0.01188	6803	0.260 ±0.061	254
Guizhou	0.226 ±0.00978	10050	0.186 ±0.047	431
Shanxi	0.222 ±0.01054	8646	0.260 ±0.057	296
Jilin	0.215 ±0.01017	9288	0.237 ±0.060	266
Jiangxi	0.207 ±0.00854	13161	0.233 ±0.053	343
Other China	0.202 ±0.00458	45805	0.216 ±0.027	1363
Heilongjiang	0.183 ±0.00850	13298	0.226 ±0.055	314
Guangxi	0.183 ±0.00632	24075	0.174 ±0.046	460
Yunnan	0.182 ±0.00859	13005	0.241 ±0.052	352
Hong Kong	0.178 ±0.00854	13170	0.241 ±0.041	585
Hebei	0.173 ±0.00768	16287	0.224 ±0.044	501
Guangdong	0.173 ±0.00154	407279	0.168 ±0.012	7097
Anhui	0.172 ±0.00794	15224	0.207 ±0.047	439
Fujian	0.171 ±0.00454	46542	0.166 ±0.031	1032
Chongqing	0.168 ±0.00643	23238	0.178 ±0.043	529
Hunan	0.164 ±0.00646	23031	0.210 ±0.040	596
Hubei	0.159 ±0.00546	32176	0.192 ±0.035	767
Outside China	0.155 ±0.00429	52069	0.215 ±0.023	1873
Tianjin	0.152 ±0.00767	16311	0.163 ±0.048	418
Henan	0.151 ±0.00636	23723	0.144 ±0.037	716
Shandong	0.145 ±0.00587	27838	0.141 ±0.034	838
Liaoning	0.141 ±0.00616	25339	0.148 ±0.038	681
Jiangsu	0.139 ±0.00413	56368	0.143 ±0.024	1619
Shaanxi	0.138 ±0.00722	18443	0.178 ±0.045	483
Sichuan	0.132 ±0.00477	42178	0.164 ±0.032	967
Zhejiang	0.129 ±0.00361	73752	0.147 ±0.023	1849
Beijing	0.120 ±0.00294	111456	0.122 ±0.015	4133
Shanghai	0.114 ±0.00310	99910	0.127 ±0.018	3001

Table 4: Overall deletion rate by province. $\delta_{uniform}$ is the deletion rate of random sample of all messages; $\delta_{sensitive}$ is the deletion rate of messages containing one of 295 known sensitive keywords. The sensitive message deletion rate has wider confidence bounds than the uniform deletion rate, but the two are correlated (Kendall's $\tau = 0.77$, Pearson $r = 0.94$).

10 About the Authors

David Bamman is a Ph.D. student in the Language Technologies Institute, School of Computer Science, Carnegie Mellon University.
Web: `http://www.cs.cmu.edu/~dbamman`
E-mail: dbamman [at] cs [dot] cmu [dot] edu

Brendan O’Connor is a Ph.D. student in the Machine Learning Department, School of Computer Science, Carnegie Mellon University.
Web: `http://brenocon.com`
E-mail: brenocon [at] cs [dot] cmu [dot] edu

Noah A. Smith is the Finmeccanica Associate Professor in the Language Technologies Institute and Machine Learning Department, School of Computer Science, Carnegie Mellon University.
Web: `http://www.cs.cmu.edu/~nasmith`
Email: nasmith [at] cs [dot] cmu [dot] edu

Acknowledgments

We thank the anonymous reviewers for helpful comments and Y.C. Sim and Justin Chiu for helpful advice about the Chinese language. The research reported in this article was supported by National Science Foundation grants IIS-0915187 and CAREER IIS-1054319 to N.A.S. and by an ARCS scholarship to D.B.

References

David Bandurski. China’s guerrilla war for the web. *Far Eastern Economic Review*, 171(6), Jul/Aug 2008.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, page 289–300, 1995.

Ben Blanchard. China tightens web screws after Xinjiang riot. *Reuters*, July 6, 2009.

Tania Branigan. Jiang Zemin death rumours spark online crackdown in China. *The Guardian*, July 6, 2011.

Laurie Burkitt. Fearing radiation, Chinese rush to buy. . . table salt? *Wall Street Journal*, March 17, 2011.

CDT. Directives from the Ministry of Truth: July 5-September 28, 2011. http://chinadigitaltimes.net/2011/10/directives-from-the-ministry-of-truth-july-5-september-28-2011/, 2011.

Royston Chan and Maxim Duncan. China sacks 3 senior officials after train crash. *Reuters*, July 25, 2011.

Loretta Chao. ‘Father’ of China’s great firewall shouted off own microblog. *Wall Street Journal*, December 20, 2010.

Loretta Chao. Sina takes aim at online rumors. *Wall Street Journal*, September 14, 2011.

Josh Chin. Following Jiang death rumors, China’s rivers go missing. *Wall Street Journal*, July 6, 2011.

Jedidiah R. Crandall, Daniel Zinn, Michael Byrd, Earl Barr, and Rich East. ConceptDoppler: a weather tracker for internet censorship. In *Proceedings of the 14th ACM conference on Computer and communications security*, CCS ’07, pages 352–365, New York, NY, USA, 2007. ACM. ISBN 978-1-59593-703-2.

Bradley Efron. *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction*. Cambridge University Press, 1st edition, September 2010. ISBN 0521192498.

Jacob Eisenstein, Brendan O’Connor, Noah A. Smith, and Eric P. Xing. A latent variable model for geographic lexical variation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, EMNLP ’10, pages 1277–1287, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics. URL `http://portal.acm.org/citation.cfm?id=1870658.1870782`.

Jacob Eisenstein, Noah A. Smith, and Eric P. Xing. Discovering sociolinguistic associations with structured sparsity. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1365–1374, Portland, Oregon, USA, June 2011. Association for Computational Linguistics.

Gady Epstein. Sina Weibo. http://www.forbes.com/global/2011/0314/features-charles-chao-twitter-fanfou-china-sina-weibo.html, March 14, 2011.

Antonio M. Espinoza and Jedidiah R. Crandall. Work-in-progress: Automated named entity extraction for tracking censorship of current events. In *USENIX Workshop on Free and Open Communications on the Internet*, 2011.

Facebook. Statement of rights and responsibilities. http://www.facebook.com/terms.php?ref=pf, April 26, 2011.

Owen Fletcher. Sina's Weibo shows strong user growth. *Wall Street Journal*, August 18, 2011.

FLOSS. How to bypass internet censorship. http://en.flossmanuals.net/_booki/bypassing-censorship/bypassing-censorship.pdf, 2011.

Benjamin Gottlieb. Ai Weiwei's release accentuated by web censorship, terse state-media. CNN, June 23, 2011.

Sandra Hernandez. The "occupy" series: Sina Weibo's new list of banned search terms. *China Digital Times*, October 21, 2011.

Human Rights Watch. "Race to the bottom." Corporate complicity in Chinese internet censorship. Technical Report 18.8, Human Rights Watch, August 2006.

Jeremiah Jenne. Ambassadors caught on tape, China edition. *The Atlantic*, February 2011.

Michael Kan. China blocks some web searches about migrant protests. *PCWorld*, June 2011.

Jo Ling Kent. Organizers call for second round of demonstrations across china. *CNN*, February 25, 2011.

Jeffrey Knockel, Jedidiah R. Crandall, and Jared Saia. Three researchers, five conjectures: An empirical analysis of tom-skype censorship and surveillance. In *USENIX Workshop on Free and Open Communications on the Internet*, 2011.

Nicholas D. Kristof. Banned in Beijing! *New York Times*, January 22, 2011a.

Nicholas D. Kristof. Blogs interrupted. *On the Ground, New York Times*, January 22, 2011b.

Brook Larmer. Where an Internet joke is not just a joke. *New York Times*, October 26, 2011.

Rebecca MacKinnon. China's censorship 2.0: How companies censor bloggers. *First Monday*, 14(2), January 25 2009.

Rebecca MacKinnon. China's "networked authoritarianism". *Journal of Democracy*, 22(2), April 2011.

Brendan O'Connor, Jacob Eisenstein, Eric P. Xing, and Noah A. Smith. Discovering demographic language variation. In *NIPS-2010 Workshop on Machine Learning and Social Computing*, 2010.

OpenNet Initiative. Internet filtering in China. http://opennet.net/sites/opennet.net/files/ONI_China_2009.pdf, June 15, 2009.

Xiao Qiang. The battle for the Chinese internet. *Journal of Democracy*, 22(2):47–61, April 2011.

Hal Roberts, Ethan Zuckerman, and John Palfrey. Circumvention landscape report: Methods, uses, and tools. Technical report, The Berkman Center for Internet & Society at Harvard University, 2009.

Andreas Sfakianakis, Elias Athanasopoulos, and Sotiris Ioannidis. Censmon: A web censorship monitor. In *USENIX Workshop on Free and Open Communications on the Internet*, 2011.

Irina Shklovski and Nalini Kotamraju. Online contribution practices in countries that engage in internet blocking and censorship. In *Proceedings of the 2011 Annual Conference on Human Factors in Computing Systems*, CHI '11, pages 1109–1118, New York, NY, USA, 2011. ACM. ISBN 978-1-4503-0228-9.

Brian Spegele. Social media helps China activists score victory for blind lawyer. *Wall Street Journal*, October 20, 2011.

David J. Spiegelhalter. Funnel plots for comparing institutional performance. *Statistics in Medicine*, 24(8):1185–1202, 2005.

Richard Sproat and Thomas Emerson. The first international Chinese word segmentation bakeoff. In *Proceedings of the Second SIGHAN Workshop on Chinese Language Processing - Volume 17*, SIGHAN '03, pages 133–143, Stroudsburg, PA, USA, 2003. Association for Computational Linguistics.

John D. Storey and Robert Tibshirani. Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences of the United States of America*, 100(16):9440, 2003.

Twitter. Tweets still must flow. http://blog.twitter.com/2012/01/tweets-still-must-flow.html, 2012.

Nart Villeneuve. Breaching trust: An analysis of surveillance and security practices on china's TOM-Skype platform. www.nartv.org/mirror/breachingtrust.pdf, 2008a.

Nart Villeneuve. Search monitor project: Toward a measure of transparency. Technical report, Citizen Lab Occasional Paper 1, June 2008b.

Wikipedia. List of blacklisted keywords in the People's Republic of China. http://en.wikipedia.org/wiki/List_of_blacklisted_keywords_in_the_People's_Republic_of_China, October 19 2011.

Michael Wines. China's censors misfire in abuse-of-power case. *New York Times*, November 17, 2010.

Benjamin Wing and Jason Baldridge. Simple supervised document geolocation with geodesic grids. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, Portland, Oregon, USA, June 2011. Association for Computational Linguistics.

Edward Wong and David Barboza. Wary of Egypt unrest, China censors web. *New York Times*, January 31, 2011.

Zhou Xin, Chris Buckley, and Sabrina Mao. Stop hoarding salt, China tells radiation-scared shoppers. *Reuters*, March 17, 2011.

Xueyang Xu, Z. Mao, and J. Halderman. Internet censorship in China: Where does the filtering occur? In Neil Spring and George Riley, editors, *Passive and Active Measurement*, volume 6579 of *Lecture Notes in Computer Science*, pages 133–142. Springer Berlin / Heidelberg, 2011.

Louis Yu, Sitaram Asur, and Bernardo A. Huberman. What trends in Chinese social media. In *Proceedings of the 5th SNA-KDD Workshop'11 (SNA-KDD'11)*, 2011.