Experiencing Censorship Emboldens Internet Users and Decreases Government Support in China*

Margaret E. Roberts†

September 17, 2015

Abstract

How do internet users react to experiences with censorship? While self-censorship of journalists and academics in China is well-documented, very little empirical research has studied how typical citizens who engage in political discussion on social media react to government censorship policies. Yet in the age of the Internet, where so much public discourse occurs online, the self-censorship patterns of these individuals have enormous implications for the spread of information and stability of authoritarian regimes. Using two social media datasets from different websites in China, I find that experience with censorship does not deter bloggers from writing about political information. Instead, in comparison to similar bloggers who were not censored, the experience of censorship motivates more political writing on more sensitive topics.

*My thanks to Jude Blanchette, Joan Cho, Jeff Frieden, Shelby Grossman, Andrew Hall, Navid Hassanpour, Will Hobbs, Haifeng Huang, Holger Kern, Gary King, James Morrison, Jennifer Pan, Elizabeth Perry, Amanda Pinkston, Susan Shirk, Daniel Smith, David Steinberg, Brandon Stewart, Chiara Superti, Dustin Tingley, Jessica Weiss and Vanessa Williamson and the participants in Harvard's Comparative Politics Workshop and UC Merced's Understanding Politics Speaker Series for helpful comments.

†Assistant Professor, Department of Political Science, University of California, San Diego, Social Sciences Building 301, 9500 Gilman Drive, #0521, La Jolla, CA 92093-0521; (617) 302-6020; meroberts@ucsd.edu, http://margaretroberts.net

Introduction

Governments frequently use intimidation as a form of political censorship and control over public discourse. Reporters without Borders estimates that so far in 2015, 159 journalists have been jailed and 20 killed worldwide.[1] In addition to traditional media, bloggers are routinely jailed around the world for their comments on social media – Reporters without Borders estimates that so far in 2015 175 netizens have also been imprisoned. Freedom House finds that over half of governments studied in 2012 had imprisoned an Internet user for content they had posted online (Kelly et al. 2012).

As many countries around the world punish their vocal critics, there is much speculation over whether these retaliatory efforts promote self-censorship in the media and the public at large. Scholars have found evidence that government intimidation can induce self-censorship in traditional media (Gomez 2000, Stern and Hassid 2012, Link 2009). However, very little empirical research has been conducted on the extent of self-censorship among typical Internet users in authoritarian contexts. Scholars have puzzled over whether censorship can still stop the spread of information in the age of the Internet (Lessig 1999, Zheng 2007, Wacker 2003, Morozov 2012, Qiang 2011); however, self-censorship of everyday online users, one of the primary theorized mechanisms through which censorship of the Internet is purported to work, has not been empirically explored in detail.

The debate about whether public speech can be controlled in the digital age speaks to larger questions about the Internet's role in the development and success of social movements and whether authoritarian states can maintain power when they have high levels of Internet penetration (Howard and Hussain 2011). The advent of the Internet has expanded the number of people who are involved in public discourse, making it more difficult for the government to target all who could speak out against the regime. If governments can successfully induce self-censorship in not just the journalistic elite but also the general public, control of information would be possible

[1] https://en.rsf.org/press-freedom-barometer-journalists-imprisoned.html?annee=2015, accessed 4/19/2015

even in the age of the Internet. If, however, awareness of censorship undermines government policies and induces more information seeking, such intimidation may counterintuitively backfire and cause further spread of information.

In this paper, I use data on censorship of social media users in China to test whether experience with censorship induces self-censorship among bloggers in China. I find, unexpectedly, that the experience of being censored does not induce self-censorship among bloggers. On the contrary, bloggers who have been censored are more likely to repost on similar topics to that which was censored, continue to post on more sensitive topics, are more likely to complain about being censored after censorship, and are more likely to be censored after censorship.

In the sections below, I begin by outlining the current literature on self-censorship of citizens in China and a theory of how experience with censorship will affect producers of information in the Internet age. I then describe the two datasets I use to better understand how Internet users react to the experience with censorship. After explaining the results of each of these tests, I conclude with implications for information freedom and authoritarian resilience in an information age.

Self-Censorship in the Age of the Internet

In China, as in many authoritarian regimes, fear and intimidation to induce self-censorship has long been a primary form of information control. Under Mao, government reprimands for the production and consumption of 'off-limits' information were extreme enough to warrant widespread fear among ordinary citizens. Anything that could be interpreted as against the state could land citizens in serious trouble. Owning a banned book or having a picture of a purged leader could result in re-education in a labor camp (Jin 1999). A neighbor overhearing you talk about your cat (which sounds like Mao) in a derogatory manner or naming your children names that sounded unpatriotic might result in government detainment (Link 2009). Extreme forms punishment for even the most mundane speech induced severe self-censorship among the general public.

After Mao, when China opened up to the world and began to reform the economy, the government ceased using many of these extreme methods of intimidation to induce self-censorship. The public sphere, particularly informal discourse, became increasingly unregulated. Whereas before even unintentional comments that could be perceived as anti-government might be severely disciplined, after reform the elite and media have ventured so far to criticize the government through various mediums including investigative journalism and public statements. As the use of the Internet expanded in China, an increasing number of citizens entered political discourse – as of 2014, China had 642 million users.[2] Criticism of the government in online forums and in social media is commonplace, and may even in some cases by encouraged by government officials to act as a pressure valve and for monitoring public opinion (Qiang 2011, King, Pan and Roberts 2013, Hassid 2012, Lorentzen 2010).

Despite increased freedom of expression over the last 20 years, fear as a form of censorship is still employed by the Chinese Communist Party, though with decreased frequency and predictability. Censorship laws in China are decidedly ambiguous: it is illegal in China to write or distribute any information online that "harms the interest of the nation", "spreads rumors or disturbs social order", "insults or defames third parties", or "jeopardizes the nation's unity".[3] Punishments are similarly ambiguous and unevenly administered – violating online information laws could results in punishment as severe as jail-time to as trivial as having your account shut down or simply removing one offending social media post. The wide range of information that could qualify under these laws keeps online users in China guessing as to what types of information are indeed off-limits and what types of punishment could be meted out for spreading the information.

Empirical evidence indicates that self-censorship of traditional media and high-profile individuals in China is common, even though the probability of any one person being arrested is quite low. Stern and Hassid (2012) estimate that about 0.2% of lawyers and journalists have been arrested, punished, or subject to state

[2] http://www.internetlivestats.com/internet-users-by-country/

[3] http://china.org.cn/government/whitepaper/2010-06/08/content_20207978.htm

violence. Even so, Stern and Hassid (2012) find many instances of self-censorship of elite in these professions, documenting hundreds of interviews in which these actors expressed fear of government reprimands and indicated that they self-censor. Link (2009) expresses concern over academic self-censorship for fear of government reprisal. Government censorship laws affect media outside of China, also – scholars have found evidence of self-censorship in Hong Kong and foreign journalists have expressed concern over their own tendency to self-censor (Lee and Lin 2006).[4]

Perhaps due to strong evidence of elite self-censorship, the academic literature has also conjectured that self-censorship may be the primary form of government control over the Internet. Kalathil and Boas (2010) argue that one of the two main prongs of Chinese control of the Internet is through promotion of self-censorship, and punitive action for those who disobey. Wacker (2003) argue that the perception of surveillance in Chinese society and a few high-profile arrests are more important than actual censorship in stopping the spread of information, and could explain why Chinese censorship efforts of the Internet have been relatively relaxed. In public discourse, self-censorship is consistently quoted as one of the primary forms of Internet censorship in China, with statements such as "Chinese society has moved into an era of self-censorship where people themselves automatically 'purify' the Internet environment."[5] However, unlike self-censorship of journalists and activists, little empirical exploration of self-censorship on the Internet exists.

Despite claims of self-censorship on the Internet, there are reasons to believe that self-censorship may be a less potent force on the Internet than within the elite. Because there are many more contributors to social media than there are journalists, lawyers, or scholars in China, the probability of any one Internet user being arrested is even lower than the number quoted above. Even though bloggers in China typically have to register their name and identity card to contribute to social media, writings on the Internet are more anonymous than those in newspapers or other forms of publication. Typical bloggers, overall, are less likely to be targeted than

[4] http://www.chinafile.com/conversation/spiked-china

[5] http://thediplomat.com/2015/03/china-self-censorship-displaces-western-threats/

their elite counterparts, as their influence over the broader population is smaller.

In addition, when intimidating the general public, the government may incur a repetitional cost that could undermine its legitimacy. Self-censorship requires that citizens be aware that the government might retaliate if they post something online. For example, when a person experiences censorship by having a blogpost removed, she knows that the government objects to its content, and that there is a certain (unknown) probability that continuing to write on that topic will make her life more difficult. Censorship, in this case, is a signal that the topic is off-limits. However, if the person opposes censorship, any anxious reaction to censorship might be overpowered by dissatisfaction with government censorship policies and might instead induce more negative views of the government, inspiring her to continue writing about the sensitive topic in order to undermine censorship. Evidence exists that even journalists can push back against the government restrictions in order to undermine censorship laws (Hassid 2010). Bloggers, who face fewer risks than these journalists may be more likely to engage in these forms of push back.

In addition, censorship is an indication that the government has something to hide, and therefore may signal to the observer of censorship that the government is weaker than they previously believed. Political scientists have found that total information blackouts may undermine government efforts, for example Hassanpour (2011) presents evidence that complete Internet blackout in Egypt during the Arab Spring undermined government legitimacy because of the censorship's observability. Indications of weakness could embolden writers to continue to post, despite the dangers of doing so (Huang 2013, Edmond 2013).

Experience with censorship reveals the topics the government would prefer not be discussed. For bloggers, knowing that particular topics are off-limits could signal a topic's importance. If the topic is one that the government deems important enough to censor, it might be worth it to continue writing about it. Whereas a protest may be uninteresting without government censorship, efforts by the government to censor the information might cause an increased spike in interest about the protest event.

Similar phenomenon has occurred with banned books, where censored books often have a wider spread than those that are technically allowed to be sold in China.[6] Censorship brings to the topics the government would rather not have discussed to the fore of the censored's mind.

Outline of Results

Self-censorship, of course, is notoriously difficult to study as it implies a lack of information rather than the existence of it. In this article, I study self-censorship by estimating the effects of Internet users' experience with censorship on their subsequent online behavior. In China, individual social posts are frequently removed by Internet content providers at the direction of government censorship directives. While in general off-limits topics in China are typically somewhat ambiguous and vary by time-period, having a blogpost censored signals to Internet users that a topics is indeed off-limits. In a high-fear environment where self-censorship is rampant, these signals should induce the users to avoid the topic, for fear of government retribution if they continue to write or read about the topic. If, however, bloggers and consumers of blogposts are not affected by the signal, we would expect that they would continue to write and read without being deterred by government reprimands.

The ideal experiment to study the effects of the experience of censorship on bloggers in China would be to censor a random subset of bloggers and compare the writings of treated bloggers after censorship to those who were not censored. Of course, such an experiment would be impossible and unethical to conduct. However, the experiment can be approximated observationally since censors miss a small subset of posts for any given off-limits topic. In this study, I locate the rare pairs of social media posts where the censors have made mistakes: where a pair of two bloggers reposted the same blogpost, but one has been censored and one was left uncensored. I study how being censored changes the sensitivity of the topics with which bloggers write by comparing the social media user who was censored to the

[6] http://blogs.wsj.com/chinarealtime/2014/10/13/rumors-of-book-ban-boosts-authors-in-china/

one who was not censored.

I find that Internet users' experience with censorship backfires against government censorship policy. I find that social media users are very likely to repost on the same topic as the one for which they were censored. Further, bloggers and other social media producers who have been censored are more likely to persist talking about the off-limits topic after censorship in comparison to their uncensored counterparts. These findings have significant implications for the efficacy of the Chinese government's censorship efforts and for authoritarian stability more generally.

Limitations

These tests have several limitations, which I hope to clarify up-front. Most significantly, I can only study how censorship affects bloggers who have already begun writing. I cannot observe how people who have never blogged are influenced by censorship. It could be that some people do not blog at all because of fear and that this type of fear has significant implications for the spread of information online. Despite this, given that millions of people already write about political topics online in China, studying the self-censorship behavior of those already participating is important in itself even if this research may not reflect every citizen in China.

In addition, these studies of self-censorship are limited to the Chinese context and to the years 2011 and 2012. Although China has one of the largest censorship programs in the world, these findings will not apply to every other country. It would be particularly interesting to study these same behaviors in more totalitarian environments and across time, particularly in the more recent periods in China under which censorship policy has been tightening. Each of these limitations could be addressed in future studies of self-censorship.

Data and Methods

I use two datasets to study the effect of censorship on bloggers' subsequent writing. For both datasets, I collect a timeline of users' posting on social media and the

censorship status of each of these posts. In the analysis section, for each dataset I will locate pairs of users who are very similar in terms of their previous censorship experience and user characteristics, and where both have reposted identical social media posts, but where one was censored and one was not. I will then examine how the experience with censorship affects the treated user's subsequent online behavior in comparison to the user who was not censored. In this section, I describe the datasets themselves in detail.

The first dataset, a longitudinal dataset of Weibo users was collected and made available by Fu, Chan and Chau (2013). Fu, Chan and Chau (2013) created a list of Weibo users with more than 1,000 followers using the User Search API, then followed these users during 2012. Their project, Weiboscope[7], provides data for 14,387,628 unique users during this time period.

The Weiboscope project collected microblog postings from each of the users' timelines in almost realtime and also revisited each users' previous posts at least once a day and frequently more than once a day to record whether the post had been removed.[8] If the post was removed, the authors documented the last time the message was seen before it was removed. They also documented the error message related to the removed post. From the authors' own experiments, "Permission Denied" indicates that the post had been censored, while "Weibo does not exist" usually indicates government censorship, but could also mean that the post had been deleted by the user themselves. While the team anonymizes the identity of the user, they include a subset of information about the user, including whether the user was "verified". Verified users on Weibo are typically those whose identity has been verified by the online platform and are typically most prominent or famous users who have more followers.

The Weiboscope data provides an almost ideal dataset to test self-censorship

[7] http://weiboscope.jmsc.hku.hk/datazip/

[8] Some posts in the dataset may be missed if they were censored before Weiboscope had time to collect them. This turns out not to be a significant issue for this project since I am comparing users instead of looking at overall censorship trends and treated and control users are collected with the same procedure. When this could affect the results, as discussed in more detail in the Discussion section, it should always attenuate the results against the main finding in the paper.

patterns because many users were followed over a relatively long time period. The approximate date and time of censorship is known, revealing the approximate time that users were "treated" with censorship. Further, we have a subset of posts that we are certain were censored, rather than voluntarily removed by the user.

However, the Weiboscope data also has some drawbacks. Because Weibo is a fundamentally interactive platform, some of the posts do not include the full context of the posts. Many of the recorded posts are short messages such as 转发微博 *zhuanfa weibo*, which simply means "reposting Weibo" and does not include the repost itself. Further, the censorship rate on Weibo is very low, only 86,083 of the total 226,841,122 messages have a "Permission Denied" message indicating that they were certainly censored. Therefore, the likelihood of finding instances of censors' mistakes is lower than it would be on other platforms with higher censorship rates.[9]

To address these challenges and provide validation of the Weibo results, I also analyze an additional dataset from another social media site. The second dataset includes a sample of 593 bloggers who were writing from September 1, 2011 to April 1, 2012 from the blogging website Baidu.com.[10] The data were collected using the data service Crimson Hexagon, which downloads the posts the instant they appear online.[11] Similar to the Weiboscope data, these data *pre-censorship*, they are recorded almost immediately after the post was written, before the government had time to censor them.

The Baidu data have advantages that the Weibo data does not. First, each blogpost is very long and complete, so there is a lot of textual information included in each of the posts. Second, the censorship rate is around 13%, much higher than the data included in Weibo, making it easier to find censors' mistakes. Even though the posts themselves were collected in realtime, the Baidu data does not have realtime

[9]The error message "Weibo does not exist" is much more common, but since it is not certain these were removed because of government censorship, I only find matches for "Permission denied" posts.

[10]The procedure for sampling these 593 bloggers is included in the Appendix.

[11]The Baidu blogging site has since been shut down by the company.

censorship information – information about which blogposts were deleted was not collected until after the full time period of blogposts was collected. For this dataset, we rely on previous empirical evidence that censors remove information typically within 1 days of bloggers' writing (King, Pan and Roberts 2013), a pattern that is also corroborated by the Weiboscope data; however, this assumption is required because we do not know the exact time each of the posts were censored.

For both datasets, we have a timeline of blogposts for a set of users, that includes censorship information for each of the users and their subsequent behavior. Both datasets coincide with a sensitive time period for the Chinese Communist Party: the purge of Bo Xilai, the Party secretary of Chongqing who was dismissed from the Party in March of 2012 on allegations of corruption and homicide. It also coincided with increases in nationalist protests and the lead up to the 2012 change of power between Hu Jintao and Xi Jinping. Therefore, both of these datasets give us a window into the reactions of bloggers after experiencing censorship during a particularly politically sensitive time period.

Matching: Weibo Users

We begin with an analysis of self-censorship within the Weibo dataset. First, I preprocess the entire dataset by removing all non-textual data from the microblogs, including emoticons and usernames. After preprocessing, I find all posts with identical text, but with different censorship statuses. I require that matches have more than 15 characters to ensure that two identical posts do not have different meanings because of their context, such as posts that only include short context-dependent phrases such as "reposting Weibo" (转发微博). To further ensure that the posts were written in the same context, I require that the matched posts were posted on the same day.

Censorship happens very quickly in China, and the Weiboscope data are no exception. The data indicate that 14% of the censored posts were not seen after they were first collected, i.e. they were deleted before the automated scraper had time to return to them. Half of the censored posts were last seen only half a day after

they were first posted and then were removed from the web. Over 80% of censored posts were last seen less than two days after they were written. However, for a few posts, censorship occurs significantly after the posts was written. Since I want to study the reaction of Weibo users to censorship and Weibo users are more likely to notice censorship for the more quickly it happens, I remove all matches where the censored post had not yet been censored more than two days after posting.

Of course, it could be that two users write similar posts but have very different histories of writing sensitive information and therefore have different propensities to be censored. To control for this, I calculate the censorship rate for each user before they wrote the matched post. I use K to K coarsened exact matching (CEM) (Iacus, King and Porro 2009) to match on the overall historical censorship rate for each user, as well as their censorship rate in the most recent time period – I match on their censorship rate 10 days before the matched post and 5 days before the matched post. I also match on the percentage of posts that went missing without a permission denied message for each user, to ensure that other less obvious forms of censorship are also identical between matched users. Thus matched users will both have experienced similar amounts of censorship overall and will have similar recent experiences with censorship. Since matched users have received similar attention in the past from censors, one is not more likely to receive undue amounts of attention from censors than others based on their past history.

In addition to censorship history, I match on a few other pieces of data available from the Weiboscope data. Since verified users, who are typically more famous and have more followers, may be more salient to censors than users who are not verified, I ensure that matched users have the same verification status. The Weiboscope data also indicates whether the Weibo post contains an image: I match on the inclusion or exclusion of an image.

With these restrictions, I find 144 matched posts, or 72 pairs of posts, which span the time period of 2012; matches appear in each month of 2012. The censored posts cover topics that we would expect to be censored during this time period. Table 1

shows the 25 words that are most highly predictive the matched posts, measured by their mutual information in comparison to a random sample of uncensored posts from the same time period.[12] This word list and a close reading of the matched posts reveal that matched posts cover a variety of topics, including posts about the protests in Shifang against the construction of paraexylene (PX) plant, post sympathetic with creating revolution similar to the Cultural Revolution, posts about protests in Hong Kong, posts mentioning the leaked online sex video showing Party official Lei Zhengfu, and many posts that talk about the removal of Bo Xilai from the CCP. Some matched posts are also complaints about censorship, from complaints about censorship of investigative journalism, to complaints about the deletion of microblogs, to complaints about the censorship of naked scenes in the screening of the film Titanic in 3-D.

The users within the matched dataset look similar to the full set of users who have experienced censorship at some point within the Weiboscope dataset, though overall there are fewer verified users among the matches. This makes some sense as the censors are probably less likely to mis-censor posts from verified users. Among users who experienced censorship in the Weiboscope dataset, 77% are male and 23% are female. This is quite similar to the matched dataset where 81% are male and 19% are female. Of the users who have experienced censorship in the Weiboscope dataset, 73% are verified and 27% are not verified. In the matched dataset there are slightly fewer verified users, with 43% are verified and 57% are not. Users within the matched dataset have been censored similar numbers of times than the full set of users in the Weiboscope dataset who have been censored at some point – within the matched dataset, users were censored a median of 11 times and a mean of 21 times over the course of the year; whereas within the set of users who had been censored in the Weiboscope data, users were censored a median of 11 times and on average 20 times over the course of the year.

Based on these similarities, we might expect that the inferences we draw from the matched population would be generalizable to the set of users who write about

[12] An explanation of mutual information is included in the Appendix.

	Word	English Translation
1	总理	Premier
2	文革	Cultural Revolution
3	重庆	Chongqing
4	总统	President
5	呼吁	Appeal
6	书记	Secretary
7	人民	the people
8	老百姓	ordinary people
9	新闻	news
10	年	year
11	政协	Chinese People's Political Consultative Committee
12	温	Wen (Jiabao)
13	集团	group
14	市长	mayor
15	处罚	penalty
16	记者	reporter
17	网络	the internet
18	报纸	newspaper
19	行政	administration
20	给了	gave
21	调查	investigation
22	犯	crime
23	人大	National People's Congress
24	革命	revolution
25	通	go through

Table 1: Words Predictive of Censorship, Matched Posts

sufficiently sensitive topics to be censored at some point, but not generalizable to the general population of Weibo users who are never censored. The similarity between the population of users who have been censored in the Weiboscope data and the matched users also give us some indication that, besides verified users, the users for which the censors made mistakes are not somehow systematically different than those who write about topics that are sensitive enough to be occasionally censored.

Do Weibo Users Persist After Censorship?

The main dependent variable we are interested in is whether the users persist in talking about the censored topic. Does out "treated group" – the censored users – take government censorship as a signal that they should avoid a topic, declining

to talk about that topic further and self-censoring? Or do they take government censorship as a signal of the topic's importance and persist talking about the topic more than their uncensored counterparts?

Because the matched pairs cover a wide range of topics, we cannot measure simply whether the treated group or control group talks more about one particular topic after censorship. What we want to measure is how *similar* are the posts they write after treatment to the matched post? If the treatment group self-censors, we would expect that they avoid the censored topic. If, however, they rebel, we would expect that they continue to write about it.

To measure similarity between the text of the censored post and the subsequent posts, I use a measure of string kernel similarity between each of the users' subsequent posts and the matched posts. This relatively straight forward measure counts the number of five overlapping characters that are common between two strings and weights this number by the length of the string. String similarity of 1 means strings are identical, string similarity of 0 means they have no overlapping words.

For example, one matched set of posts contained the following text about Lei Zhengfu, the Party official who was caught on video with an 18-year old woman:

"【重庆雷政富同志，请辟谣！】资深调查记者举报，重庆市北碚区委书记雷政富（正厅级），与重庆市开县赵家镇18岁女青年赵红霞发生不正当男女关系后，又动用权力将赵红霞抓捕，试图封口。由于本人水平有限，无法识别图片真伪，群众的眼睛是雪亮的，诚邀广大网友一起鉴定！"

"Chongqing Comrade Lei Zhengfu, please spread rumor! A senior investigative reporter reported that the Chongqing Beibei District Secretary Lei Zhengfu (at the department level), had an improper relationship with an Chongqing city, Kaixian county 18 year old woman Zhao Hongxia and used his power to arrest Zhao Hongxia to try to seal it. One person's ability is limited, can't validate the authenticity of the picture, everyone's eyes are good, I invite all the Internet users to try together!"

The matched dataset contains two authors who both had reposted this message, one of whom was censored and the other was not. Two days later, the censored user writes a similar post:

"【声明】重庆北碚区委书记雷政富"淫秽视频"事件，有人邀请大鹏前往"澄清"，毫无必要：第一，大鹏不是本消息的首发者，也未在微博中断言，仅起到推动作用，让大家一起关注、鉴别；第二，大鹏不是纪委领导也不是宣传领导，记者同志也已前往，你们向他和公众交代就可以了。继续欢迎新闻爆料"

"Statement: The obscene video scandal involving the Secretary of Beibei District in Chongqing, Lei Zhengfu, someone asked Da Peng to clarify, this is unnecessary. First, Da Peng is not the first one who break this story, and did not make any conclusion, only spread the story to the public, so everyone can pay attention, and to evaluate/discern. Second, Da Peng is not an official for the Commission for Discipline Inspection, nor an official for the Propaganda Department. A reporter has left (to find more about the story), you guys only need to respond to him and the public. Continue to welcome other breaking news."

These two texts have many overlapping strings included in the long string Chongqing Beibei district secretary Lei Zhengfu (重庆北碚区委书记雷政富). As a result, the two texts have high string kernel similarity of 0.25. Since words in Chinese are typically two characters long, and rarely more than four characters long, in texts as short in length as the Weibo data even one 5-character overlapping string is an indication that two posts are very similar in subject matter.

For each user in the matched dataset, I measure string kernel similarity between the matched posts and each post that the user wrote 10 days before and 10 days after the match. Figure 1 plots average string kernel similarity for treated and control users by time before and after the post using a smooth local fit.[13] As you expect,

[13]The matched post is excluded from the time plot.

string kernel similarity is the highest for both groups right around the time period of the matched post. However, surprisingly, while on average the treated and control groups talk similarly about the matched topic before censorship, after censorship the treated group writes significantly more similarly to the matched post than the control group. Being censored, at first glance, seems to inspire more *similar* writing to the censored topic than a trend away from that topic.

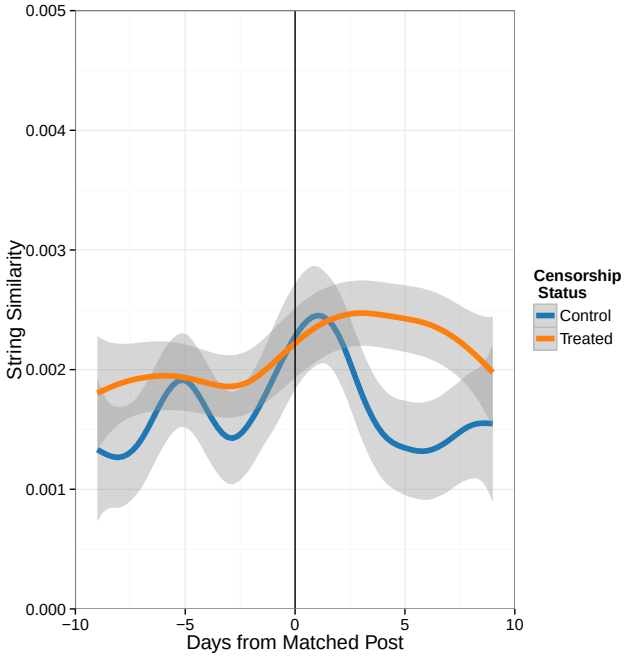


Figure 1: String Similarity Before and After Matched Post

String kernel similarity is a useful metric in this context because high levels of string kernel similarity for such short posts like microblogs are also typically quite similar to each other in terms of topical content. However, it could be that the similiarity that we are measuring is not related to the topic of the matched post, but rather is picking up similarity in the ancillary words within the post. To make sure that we are measuring similarity to the *sensitive* content of the matched post, I take the 25 words most related to the matched posts in comparison to a sample of uncensored posts from Table 1. For each post the users wrote in the 10 days before and after censorship, I measure the number of times the user mentions a word within this list and divide by the post length to standardize across posts.

Figure 2 plots the average percentage of each post that is one of these 25 words. While on average the treated and control group use the words similarly before the matched post, the treated group is much more likely to use these words after censorship. In other words, the treated group is more likely to use the word that predicts censorship *after* being censored than the control group, even though they have received a signal that these words are off-limits.

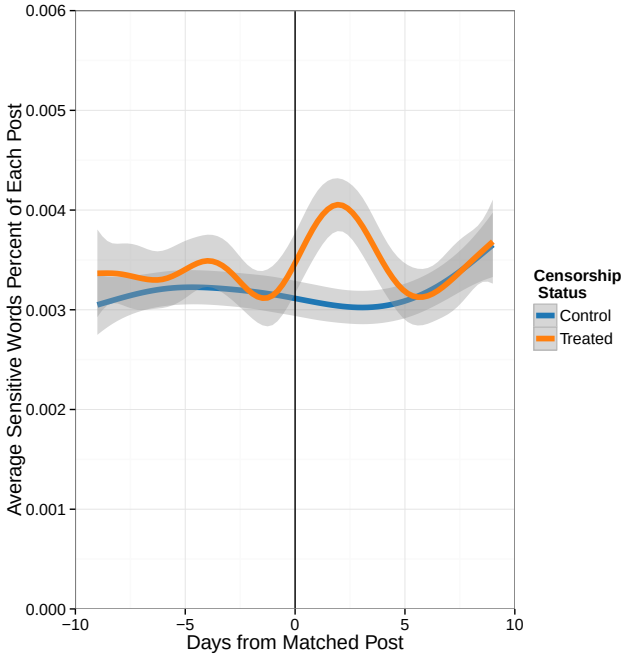


Figure 2: Sensitive Words in Posts Before and After the Match

To further validate that the differences between the treated and control groups after treatment are not an artifact of the method, I decided to look more closely at the days on which the treated and control groups diverge. I took a random sample of posts from the second, third and fourth days after censorship, stratified by the matched pair. Without looking at treatment status, a research assistant coded each of 3,000 posts on a scale of 1 to 10 of being 1 not similar at all to the matched post, and 10 very similar to the matched post.

Shown in Figure 12 in the Appendix, the human coded measure of similarity is very correlated with both string kernel similarity and the counts of the sensitive words, which provides confidence that string kernel similarity and counts of sensitive

words are indeed measuring content similarity. Like in the previous analyses, there is a strong, significant relationship between similarity and treatment in the days following censorship. On average, the censored group had a similarity metric of 1.17, while the control group had an average similarity metric of 1.12, a difference that is significant at the .05 level. Figure 3 shows the relationship between the human coded similarity and treatment – around 75% of the posts that were similar to the matched post in the 2 to 4 day time period were written by the treated group.

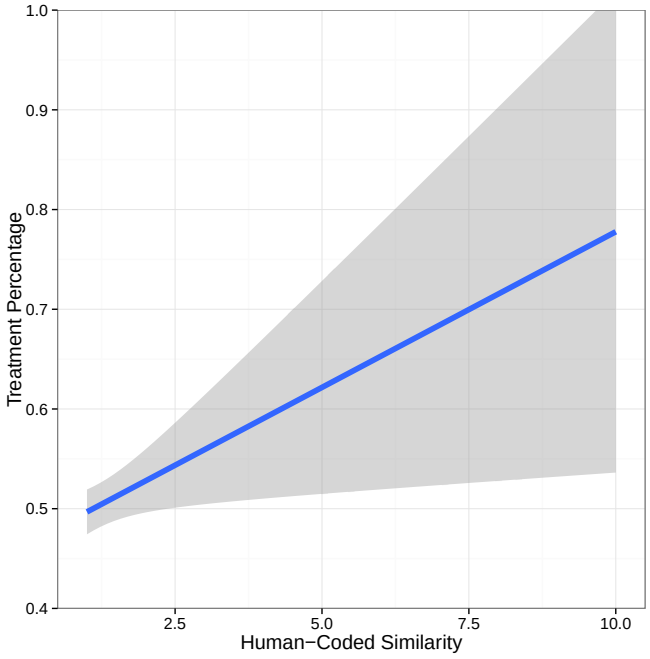


Figure 3: Relationship Between Human Coded Similarity and Treatment

It could be that the treated group continues to talk about the same topic as the post by which they experienced censorship in order to repent for their wrongdoings, or take a position on the topic that is more in line with government policy. Perhaps netizens think that if they write a post *supporting* the government's position on the same topic on which they were censored, then they will redeem themselves in the eyes of the government. This would be consistent with our results, but would have very different implications for censorship.

To test this, we code each of the 3,000 blogposts on days 2, 3 and 4 after censorship with whether or not they are critical of the government. Shown in Figure

4, we find in fact that for high similarity posts, treated bloggers are in fact *more* critical than control bloggers. For high similarity posts, almost 50% of the posts are explicitly critical in the treated group, whereas about 20% of the high similarity posts are critical in the control group. In sum, not only to treated bloggers seem to talk more about censored topic after treatment, they also seem to be more critical than their control counterparts.

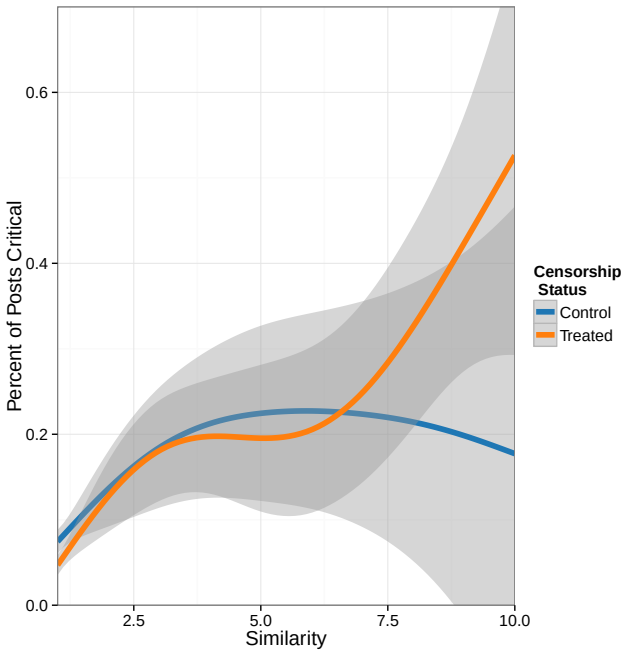


Figure 4: Treated Bloggers are More Critical for Similar Posts than Control Bloggers

Last, we code each of the 3,000 blogposts on days 2, 3, and 4 as whether they contain complaints about censorship. Treated bloggers were twice as likely to complain about censorship after the match than control bloggers – 1 in 50 of the posts of the treated bloggers complained explicitly about censorship, whereas only 1 in 100 of the control bloggers complained about censorship online (a t-test between these differences has a p-value of 0.054). This provides evidence that the treated bloggers did indeed notice the censorship, as they actually talk about that experience in their subsequent posts. It also indicates that they feel more rather than less empowered to object to censorship directly to their censors after experiencing censorship.

Even though the treated and control groups have identical censorship rates before the matched posts, the treated group is also more likely to be censored in comparison to the control group after censorship, see Figure 5. This same pattern occurs for posts with the censorship status "Weibo Does Not Exist", as shown in Figure 6. While part of this effect may be due to the topical persistence of the treated group, these differences in censorship after matching are too stark to be completely explained by the fact that the censored group tends to continue talking about the topic more than the control group. We expect that the differences in censorship rates are partly due to increased attention by the censors, who we know to flag users after censorship. This makes the results even more striking, since users are not only persisting after being censored, but are persisting in talking about the same topic and complaining about censorship in the face of increased scrutiny by the censors.

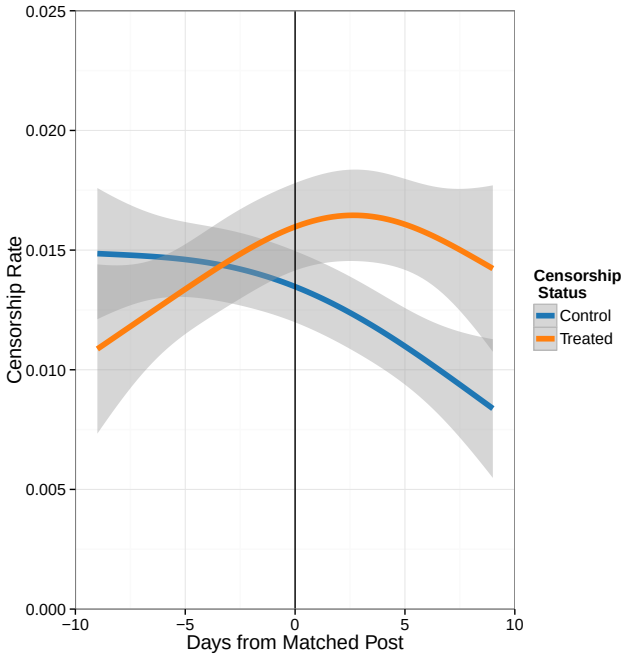


Figure 5: Censorship "Permission Denied" Rates Before and After Matched Posts

Case Studies

Why would the treated group talk *more* about the censored topic when they have just received a signal that the topic is off-limits and are under heightened scrutiny

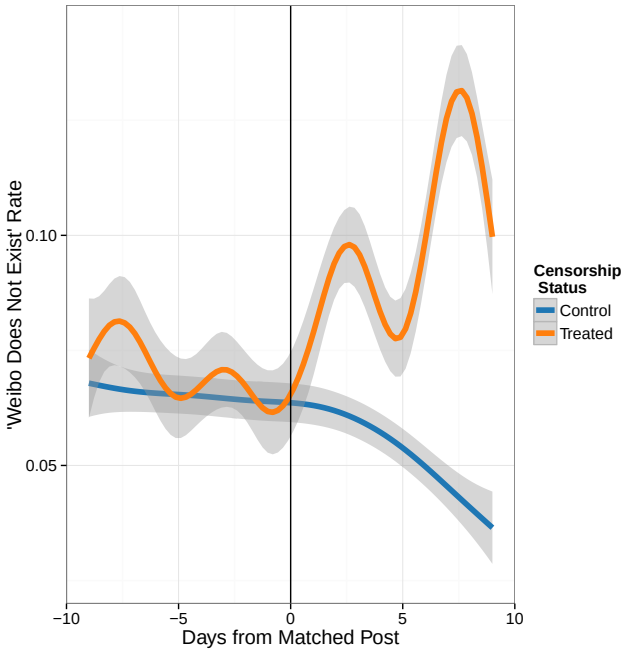


Figure 6: Censorship "Weibo Does Not Exist" Rates Before and After Matched Posts

by the censors, who have been alerted to their bad behavior? A closer look at a few treated units provides some insight into the thought process of the treated users. Take the User Zhang, who frequently blogs about corruption and the rule of law in China. User Zhang wrote a post "voting" for the decrease in censorship of the Internet, which, ironically, was censored. Whereas User Zhang only talked about the Internet four times in the 10 days before the post, User Zhang has 12 posts in the 10 days after censorship that talk about the Internet, 6 of which occur the day after User Zhang was censored.

Why would User Zhang be so relentless? User Zhang opposes censorship so strongly, that he will do everything to defy the censors. He sees censorship as indication that the government is trying to cover up corruption – he sees censorship as the direct result of corrupt officials. In the few days after being censored, User Zhang writes:

"//@经济抠门男：这是贪官们怕微博把他们都给爆料了吧，想挡住大伙的

嘴，徒劳”

“It’s because the corrupted officials are worried about Weibo will spill all their (negative) secrets, so they try to shut everyone up, it’s useless.”

“网络反腐，今天我们败了，但明天我们一定会迎来胜利。都贪污，你不贪，你就没办法在官场混！”

“Online anti-corruption, although we lost today, but we will have the victory tomorrow. Everyone is corrupted, if you don’t, you can’t survive in the government.”

Of course, User Zhang’s match User Liu also frequently writes about corruption and the rule of law in China. User Liu also opposes censorship, having written an identical post to User Zhang “voting” for a decrease in censorship on the Internet. However, User Liu is not reminded of how much he hates censorship because, unlike User Zhang, his complaint was not censored. User Liu therefore writes fewer posts opposing censorship and even about the Internet in the time period directly following the match.

These examples of complaints about censorship occur not only when the matched post is about censorship itself, but also in other contexts. A close reading of User Zhang and User Liu’s posts indicate that these two microbloggers are somewhat “right” in ideology in China, i.e. their views are pro-Western and support human rights. They also do not express support for Bo Xilai or for Mao. User Zhu, on the other hand, we would think of as somewhat “left” in China. He regularly reposts supportive comments about Mao, supports socialism, and is skeptical of the West.

User Zhu is censored when he writes:

“共产党领导下的社会主义国家不唱红，难道要唱黄唱黑？”

“The Communist leaders of the socialist country don’t sing red, don’t tell me

that they sing yellow and sing black?"

Bo Xilai, the mayor of Chongqing who was removed from the Party in 2012, supported "singing red", or supporting Maoist ideas. User Zhu is complaining that the Party itself doesn't support the Communist revolution, but rather "sings black and sings gold", or is corrupt.

Despite being censored, User Zhu writes more about Mao Zedong after censorship than before, outpacing his uncensored Match User Li. In fact, the most similar post in terms of string kernel similarity to the matched post in the 20 days surrounding the match from either user is a post written by User Zhu *after* censorship, again comparing the Communist Party officials unfavorably to officials who served under Mao.

Why is User Zhu so persistent? In User Zhu's own words, written after he was censored:

"网民 都 知道 ，虚拟的网络存在许多问题 ，但是它目前绝对是曝光贪官污吏的主要利器！对这一利器，政府应该扶持、鼓励 ，而不要去压， 甚至打击，否则，会让人民误解为你们想故意使利器变钝、生锈，这与以习主席为首的党中央反腐败行动背道而驰！你说是不是？ "

"Internet users all know, the Internet has lots of problems, but right now it definitely is the main tool to expose corruption officials! This is a weapon, the government should support and encourage it, rather than pressuring or even fighting it, otherwise people will misunderstand that you all (the government) want to blunt this weapon, and make it rusty. This (idea) contrasts the anti-corruption movement leads by Xi and other central party leaders. Am I right?"

In essence, the act of censorship has been interpreted by this user as weakness – and the censorship system itself indicates that even the top leadership has something to hide. Before the matched post, User Zhu's censorship rate was only 13%, which doubled to 28% in the ten days after the matched post. Despite being flagged by the censors, User Zhu persists, continuing to talk about the political topics he believes

are important and criticizing censorship itself.

Matching: Baidu Users

While the Weiboscope data is much larger and provides more potential matches for estimating the effects of censorship, the 593 Baidu bloggers generally write longer, more complete blogs and therefore provide a good counterpoint to the fast-paced, short Weibo data. This additional dataset also provides an opportunity to study the same phenomenon on another platform, and ensure that blogger persistence after censorship is not only something that occurs on Weibo. In this analysis, because I have relatively fewer posts, I try to leverage the length of the text as opposed to the frequency of the text in understanding the effects of censorship on subsequent writing.

The Weiboscope data are very short and identifying identical blogposts is quite straightforward, since very little extra text is added to matching blogposts. The Baidu blogs are much longer, which mean that more small changes are made to reposts. Further, unlike the Weibo data where small changes can constitute half of the post, small changes in the Baidu blogs data are often very small edits to the post. Many of these changes do not substantially affect the content of the post or its propensity for censorship, and therefore throwing out matches that are not perfectly identical would lead to a significant loss of power.

In order to avoid discarding substantively similar, but not perfectly identical posts, I use an algorithm for text matching developed in Roberts, Stewart and Nielsen (2015) to identify matches. Roberts, Stewart and Nielsen (2015) develops and algorithm called Topical Inverse Regression Matching (TIRM), which matches on 1) the topics estimated in the posts and 2) the estimated probability that the post will be treated. In this context, this means we will first match on the topics of the posts so that the amount the post talks about each of 200 topics is almost identical in matched posts. Second, we will ensure that the estimated probability that a post is censored is the same between matched posts, which ensures that if any text is different between the two matched posts, it is unlikely to be related to censorship.

For example, if the author name of the post is included and differs between the two posts, this will unlikely affect whether or not the post is censored and so the two posts will be allowed to be matched despite this difference. However, if a sentence is added that talks about a protest in one post, but not in another, then these posts will not be able to be matched because their estimated probability of treatment will differ substantially. More details of the algorithm can be found in Roberts, Stewart and Nielsen (2015).

Besides matching on the text of the post, I match on the date of the post, as posts may be more sensitive during the time period of particular events. Like in the previous analysis, I also match on the previous censorship experience of each blogger. However, because the Baidu data has many fewer blogs per user than in the Weiboscope data (on average 10 per user, rather than around 4,000 per user in the Weiboscope data), the previous censorship experience number is quite noisy. To ensure that bloggers write about similarly sensitive topics before censorship, I leverage the long textual data contained in each of these posts, which allows me to estimate the previous sensitivity of each of the blogger's posts based on the words in them. I match on this measure in addition to the previous censorship rate. Since, unlike Weiboscope, I cannot tell with complete certainty whether the missing blogs are due to censorship or to personal deletion of posts, I also require that matched posts have a greater than 0.1 predicted probability of censorship, as those posts are most likely to have been taken down by the censors rather than deleted by the bloggers themselves and thus provide the hardest test of the theory.

Balance After Matching

The matched posts must be extremely similar in order to provide good counterfactuals of one another. I enforce an extremely high standard of similarity: matched posts must be written within two weeks of each other. I create bins for topics sin TIRM matching the requires posts to be identical or very close to identical; they are typically both reprints of the same post. This ensures that idiosyncrasies within the post did not cause difference in censorship: the censors must have simply missed

the post or failed to censor it.

Of course, since the censors are very precise, there are very few posts that are essentially identical and written within similar time periods that do not have the same censorship status. Out of the almost 40,000 total posts with which I began, I found 46 posts that fit this criteria, 23 treated and 23 control. While small, the precision of the matches give us the leverage to see how differential censorship likely based on mistakes influences the subsequent posting of bloggers.

The posts and the context within which the posts were written is quite similar between treated and control groups. Most matched posts were written within three days of each other. There is no significant difference between treated and control dates, and the dates of the censored blogposts do not come systematically before or after their uncensored matches. The matched bloggers were very similar also, the difference in historical censorship rates between treated and control bloggers is not statistically significant (p-value .31). TIRM estimates that the treated and control bloggers wrote equally sensitive posts before the match.

The censored posts themselves are for all practical purposes identical – reprints of the same blogpost with little variation between them. Because I am not matching identical posts like in the Weiboscope data, to make sure that the matched posts are sufficiently similar, I calculated a string kernel similarity between all matched posts. As shown in Figure 7, the matched posts are nearly identical, with over 95% of string kernels matching in any pair.

The posts are similar in content to the matched posts in the Weiboscope data – they cover topics such as the revival of the Cultural Revolution, ousted leader Bo Xilai, complaints about censorship, reports on the arrests of dissidents, and reports on protests, including a report of a crackdown of a group of people gathered in the streets to commemorate Mao. A list of 25 words highly associated with the matched post in comparison to a random sample of uncensored post is shown in Table 2. From a qualitative read of the posts, all would be considered sensitive by the Chinese government and all fit previous findings of what the Chinese government

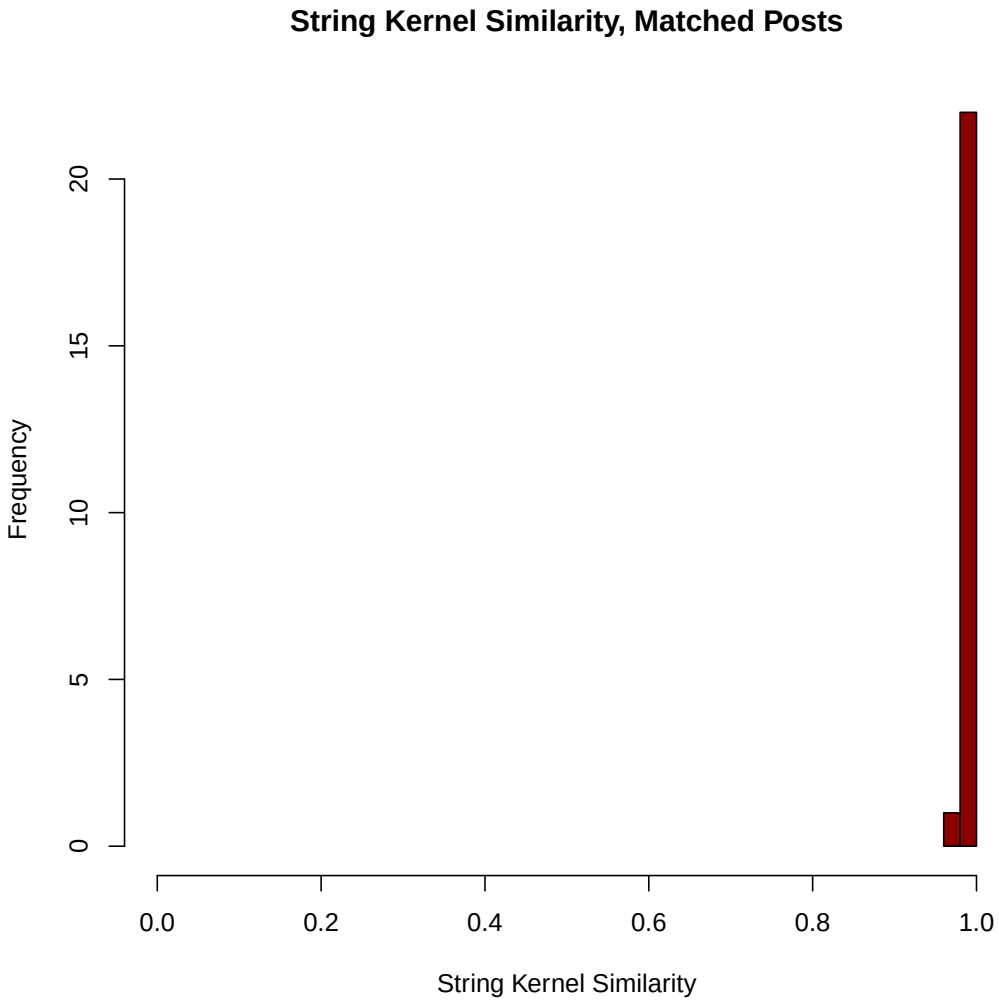


Figure 7: String Kernel Similarity Between Matched Posts

censors.

Results

Having identified uncensored posts that are highly similar to those that were censored, I now estimate the effect of experiencing censorship on the subsequent writings of these bloggers. For each matched blogger, I find the five posts they wrote before and after being censored.[14] Like the previous analysis, the question of interest is: are censored bloggers more or less likely than those who were not censored to post about the topic of the censored blogpost? If bloggers avoid the same topic of the

[14]I use posts as opposed to days that were used in the Weiboscope analysis because bloggers do not post every day, like they do on Weibo.

	Word	Translation
1	干部	cadres
2	中国	China
3	国家	country
4	毛泽东	Mao Zedong
5	社会主义	socialism
6	提出	propose
7	建设	build
8	打黑	fight corruption
9	官员	official
10	贪污	corruption
11	领导	leader
12	共产党	Communist party
13	调查	investigate
14	分类	classify
15	默认	default
16	中纪委	Central Commission for Discipline Inspection
17	唱红	sing red
18	薄熙来	Bo Xilai
19	政府	government
20	社会	society
21	群众	crowd, the masses
22	邓小平	Deng Xiaoping
23	势力	power
24	认为	think
25	改革	reform

Table 2: Words Predictive of Censorship, Baidu Matched Posts

post on which they were censored and avoid other sensitive topics then we would expect that they are self-censoring compared to their uncensored counterparts. If they post on the same topic and more sensitive topics, their experience with censorship may have made them more defiant against the government.

Like the Weiboscope data, in Figure 8, I estimate the string kernel similarity between the matched post and each of the post written before and after treatment. While there is no difference between string kernel similarity before censorship, after treatment, string kernel similarity increases within the treated group, which string kernel similarity decreases in the control group. While we do not quite have enough power to distinguish the trend between treatment and control, the direction of the effect is consistent with the Weiboscope findings.

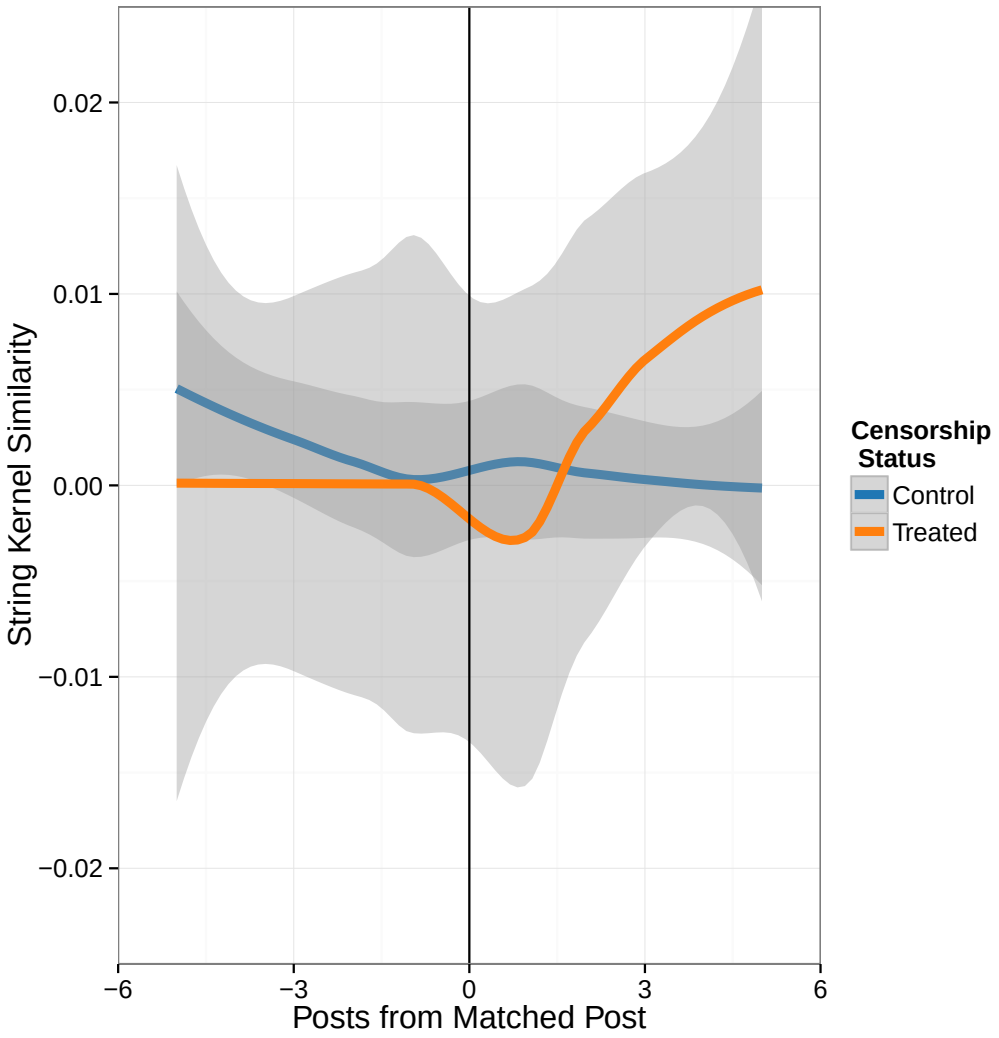


Figure 8: The Effect of Censorship on Similarity in Post-Censorship Writings

Because the Baidu data contain such long posts, we can estimate the probability of censorship for each post using TIRM. For each post before and after the matched post, we estimate the sensitivity of the post by TIRM's estimated probability that each of these posts will be censored. As shown in Figure 9, while there is no difference in sensitivity between treated and control before the matched post, the estimated sensitivity of the subsequent posts for the treated group are much higher than the estimated sensitivity in the control group.

Third, I looked at the probability of censorship for the posts after treatment for treated and control bloggers. Shown in Figure 10, treated bloggers were much more likely to be censored after treatment than control bloggers. As in the Weiboscope

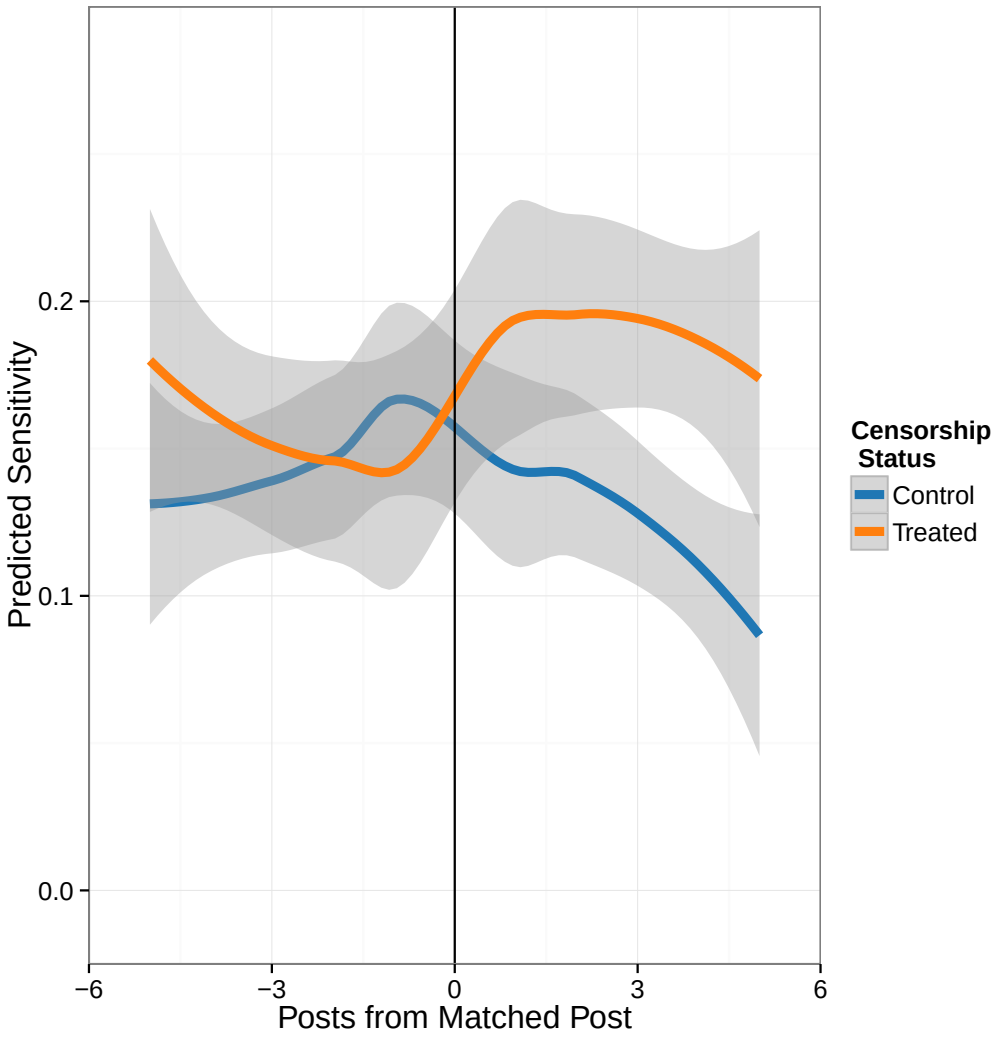


Figure 9: The Effect of Censorship on Sensitivity

data, the difference in censorship between treated and control after the matched post is much starker than the estimated sensitivity would suggest, which may mean that the censored bloggers are flagged after censorship.

A closer examination of the topical content of the posts reveals that the topics that the treated bloggers switch into after censorship that make their posts more sensitive are nationalistic topics that invoke Chinese history to criticize the current regime, call the current regime a "traitor" to the Chinese people and complain about censorship. Figure 11 shows how the topic proportions in "traitor" Topic 1 with probable words such as "the Chinese people, disaster, history, thousands, opium, extreme" and "traitor" Topic 2 with probable words such as "people,

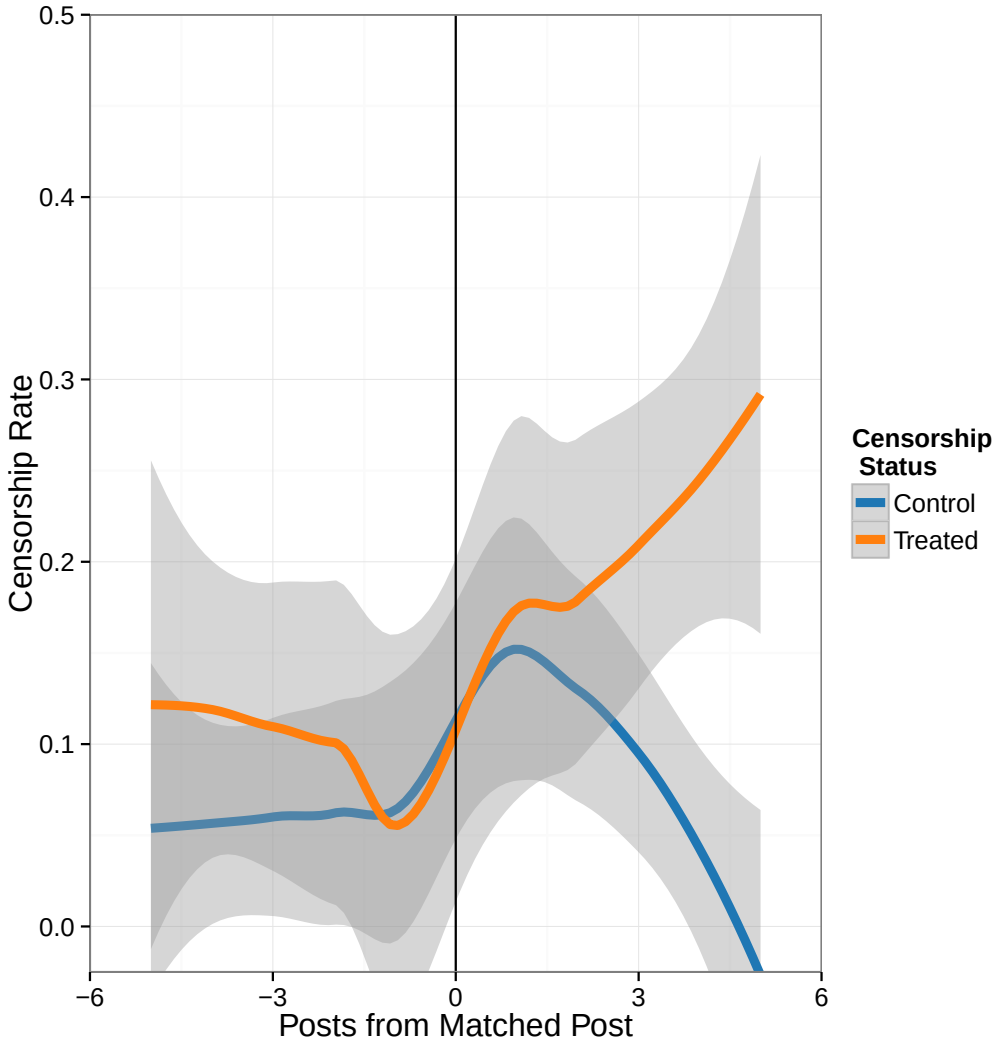


Figure 10: The Effect of Censorship on Subsequent Censorship

traitors, Republic, nationality, support, dedication, Communists" change in response to censorship.

A close reading of blogs within these topics indicates that these posts have heightened criticism of the regime. An excerpt from a post that loads onto Traitor Topic 2, written by a treated blogger after censorship has intense criticism of a regime the blogger views as having betrayed her:

"共产党已经背叛了共产，蜕变为私产党，廉洁的党已经背叛了廉洁， 衰败为肮脏的党， 公正的党已经背叛了公正， 腐化为淫邪的党，诚实的党已经背叛了诚实， 变换为虚伪的党，光荣的党已经背叛了光荣，退化为无耻的党 ，伟大的党已

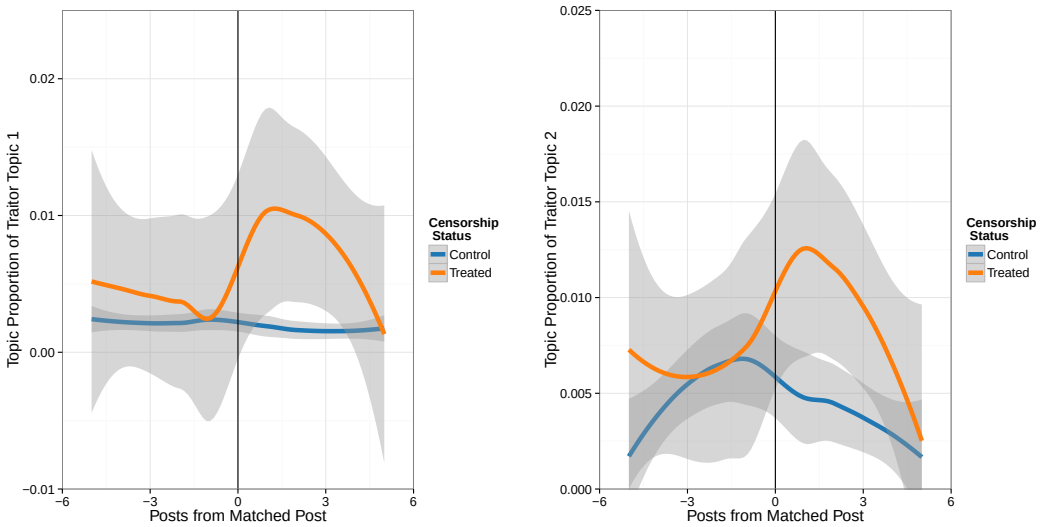


Figure 11: The Effect of Censorship on "Traitor Topics"

经背叛的伟大，沦落为渺小的党，正确的党已经背叛了正确，沉沦为恶劣的党，人民的党已经背叛人民，堕落为少数人党 。"

"The Communist Party has betrayed Communism, has disintegrated into a private property Party, the honest party has betrayed honesty, has become a dirty party, the fair party has betrayed fairness, has become a corrupt and excessive party, honorable party has betrayed honor, transformed into a hypocritical party. The glorious party has betrayed glory, has become shameless, the great party has betrayed greatness, has become an insignificant party, the correct party has betrayed correctness, has become a wicked party, the People's party has betrayed the people, has become only a small number of people's party."

An excerpt from a post that loads onto Traitor Topic 1, written by a treated blogger after censorship has intense criticism of censorship:

"大辩论正在成为现代人权的基本内容；可是在中国，随着对大民主的声讨...社会大众越来越被排除在民主之外 ，民主完全变成了精英集团内部的一种闺房游戏 ，只有在公众视野之外才能尽情嬉戏 。 未来历史学家将无论如何也无法理解这一代中国人的高超智商： 把允许自由看作是极左专制 ， 把禁止自由看作是伟大民主 。"

"Debate is an essential element of modern human rights; but in China, while democracy has been condemned, society has increasingly been excluded from democracy, democracy has completely changed into an elite haram game, which only can be played outside of the public's eye. In the future, historians will have no way to understand this generation of Chinese people's high IQ: to permit freedom is seen as ultra-left, to prohibit freedom is seen as democratic."

Discussion

Why does experience with censorship embolden bloggers and cause them to complain about censorship, despite increased scrutiny from government censors and indications that their writings are off-limits? Clearly for the producers and consumers of online media studied in this paper, the small increase in probability of government reprisal was outweighed by countervailing forces that embolden Internet users to persist in speaking about the topic that initially interested them and to write more about politically sensitive information in China. However, what countervailing forces inspired this reaction is subject to speculation.

There are many reasons why experience with censorship might backfire against government censorship policies, and future work could help unravel these mechanisms. From the examples discussed in this paper, it seems that bloggers are indeed angry with their censorship experience and they must persist in their writing as a form of protest. This suggests that bloggers could be pushing the limits of censorship because they disagree with it and hope to undermine it.

It could also be that censorship acts as a signal of topical interest or a badge of honor. Censorship does not only remove the blogpost, it also signals what the government thinks is important, or is scared of. Perhaps bloggers hope by continuing to write on the topic they will gain a greater following or more interest in their blog because this topic will prove to be important in the future. Censorship could be a badge of honor for a blogger who is trying to gain legitimacy in a competitive

information environment.

Even more subtly, it could be that experience with censorship creates more certainty about what is censored and therefore creates less fear among citizens in China. Experience with censorship may not be as bad as the producers or consumers of blogposts had expected it might be, reducing their propensity to self-censor. This mechanism would be consistent with Stern and Hassid (2012) and Link (2009) who find that the ambiguous nature of censorship among the primary motivations for self-censorship in China. If this were the case, self-censorship would counter-intuitively only take hold when the government failed to follow through, when the fear of the unknown is more scary than that which they have already experienced.

Of course, these empirical results do not mean that self-censorship does not exist among typical citizens in China. Citizens who do not blog or those who never blog on political topics may have already self-selected out of political discourse because of self-censorship. It could be that self-censorship does not affect bloggers who have already decided to join the political discussion in China, and that experience with censorship would increase self-censorship propensity of those who have already opted out.

Still, for the large subset of people who blog on political topics, experience with censorship does not seem to deter them from their task at hand. And the fact that they are not deterred could have significant implications for the spread of online information in China. Instead of discouraging users to continue to post, the government is instead encouraging more rebellion. In certain circumstances, the reputational cost of censorship may outweigh its efficacy in stopping the spread of information.

What is clear is that current scholarly and public perceptions that self-censorship is the main force of online control in China have little support based on the evidence within this study. Widespread self-censorship may not extend to political bloggers who write about political topics in their spare time. For authoritarian regimes, fear-mongering in the age of the Internet may be more bark than bite, and as a

consequence backfire by creating worse perceptions of government and inducing the further spread of information.

Conclusion

In this paper, I provide empirical tests for self-censorship of Internet users in China. Using two unique datasets, I find that bloggers who experienced censorship react negatively against the government and are more likely to write about similar topics to the one that was censored and other more sensitive topics. These empirical tests indicate that self-censorship may not be as widespread as previously thought online in China and that the observation of censorship may backfire against government censorship laws.

Understanding the underpinnings of self-censorship is imperative to understanding the efficacy of government censorship in the age of the Internet. If typical Internet users can be controlled through fear of repercussions, then we should expect governments to expand the observability of censorship to signal to users what topics are off limits. If, however, such observable censorship backfires, governments may have less control over the Internet environment that previously supposed. In this case, unobservable censorship that increases costs of access may be the more pernicious threat to the spread of information.

Future work on self-censorship of Internet users could be expanded to other political actors and other countries. Is self-censorship more common in high-profile bloggers than the typical bloggers studied here? How does awareness of censorship affect the sharing of information online, outside of reading? Do these results hold in other countries? Understanding the comparative aspect of self-censorship, both within countries and cross-nationally will help unpack how government control of the Internet affects the spread of information to people all over the world.

Appendix

Blogger Sampling Procedure

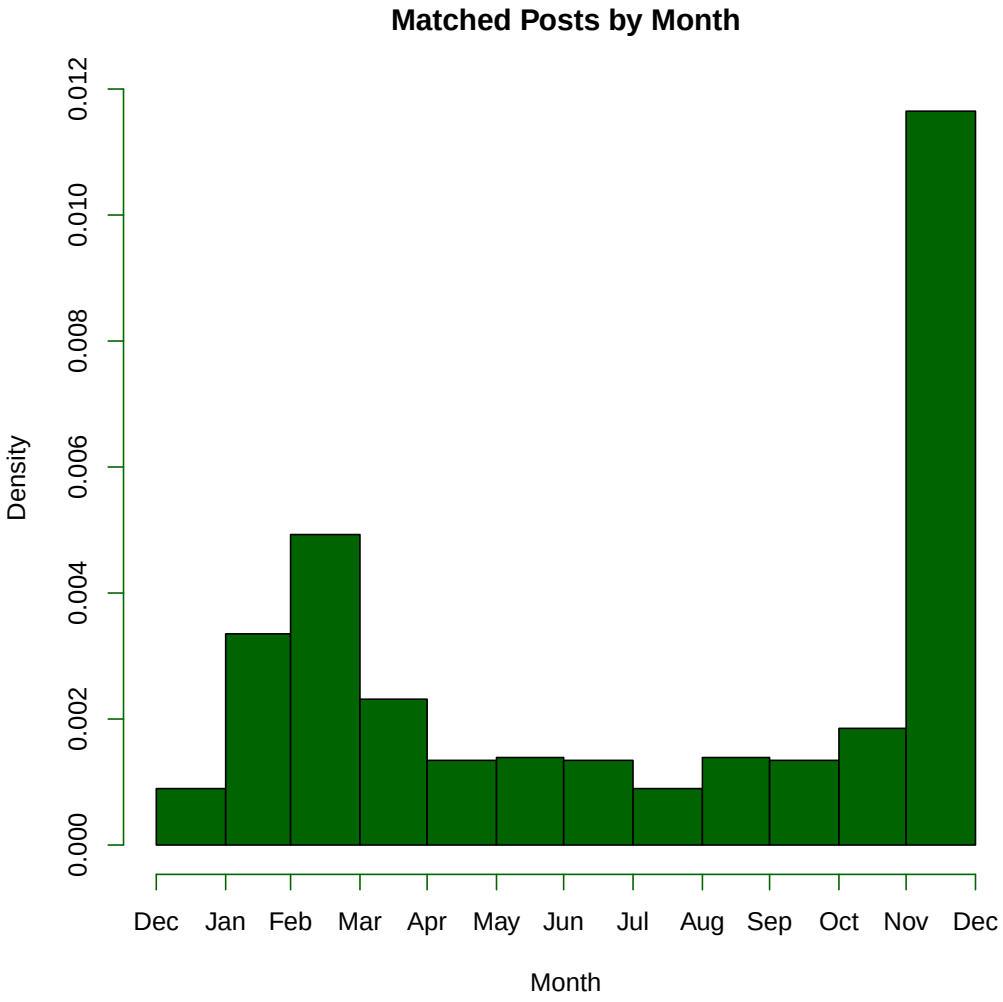
Social media posts from Crimson Hexagon must be sampled by keywords. In order to identify individual bloggers talking about current events and not spam, a list of keywords was identified on a wide range of ongoing topics in China: some political, some economic, and some cultural. A sample of bloggers were taken from all bloggers on Baidu that had blogged about one of the following keywords. Then for each of these bloggers, the remainder of their blogposts were downloaded from the time period.

Translated Keywords:

drought, AIDS, angry youth, Baidu copyright, Bo Xilai, labor union, corruption, currency, death penalty, environment, Environment Protection Agency, food prices, food safety, hacker, Guo Meimei, Huang Yibo, immigration, inflation, Iran, nuclear weapons, Japanese earthquake, Kim Jong Il, Kissinger, Kungfu Panda, lead acid batteries , Libya, microblogs, nuclear power plant, one child policy, Osama bin Laden, Pakistan, PLA, power prices, quantitative easing, rare earth metals, real estate tax, second rich generation, solar power, South China Sea, State information office, Su Zizi, fifty cent party, Three Gorges Dam, Tibet, Wen Jiabao, political reform, Wu Bangguo, Wuhan rapist, Xi Jinping, Yao Jiaxin, milk powder scandal, Zimbabwe, Qingyang bus incident, Ai Weiwei, Inner Mongolia, Boxun, censorship, Chen Guangcheng, democracy, Jasmine revolution, Falungong, Fuzhou bombings, Google, Green Dam software, labor strikes, Li Chengpeng, Lichuan protests, Liu Xiaobo, Mass incidents, princelings, Qian Yuhui, social unrest, Syria, Taiwan, Tiananmen protest, Uyghur protest, Zengcheng protest, John Hunstman, Africa investment, Doupo Canqiong, Da Renxiu, Groupon, Tennis star Lina, Track star Liu Xiang, Cooking, Disney theme park, Education reform, health care reform, indoor smoking ban, Let Bullets Fly, Peking opera, social security, space shuttle endeavor, traffic in Beijing, TV show Meirenxinji, Video game Saierhao, World Cup

Matched Posts By Month

This histogram shows the matches per month for the Weiboscope data. Matches occur in each month of 2012, but are concentrated in March and December of 2012.



Mutual Information

Mutual information identifies words that distinguish between two corpuses. In both the Weiboscope and Baidu data, I use words with high mutual information to describe how the matched post was different from the general group of uncensored posts. Words with high mutual information that are related to the matched posts are particularly sensitive, or related to censorship.

For each word, mutual information is a measure of how much that word could help us to predict whether the document was from the matched corpus or the randomly selected unmatched set. For a complete description of how to calculate mutual information, see Manning et al. (2008).

Relationship Betweeen String Kernel Similarity, Sensitive Words, and Human-Coded Similarity

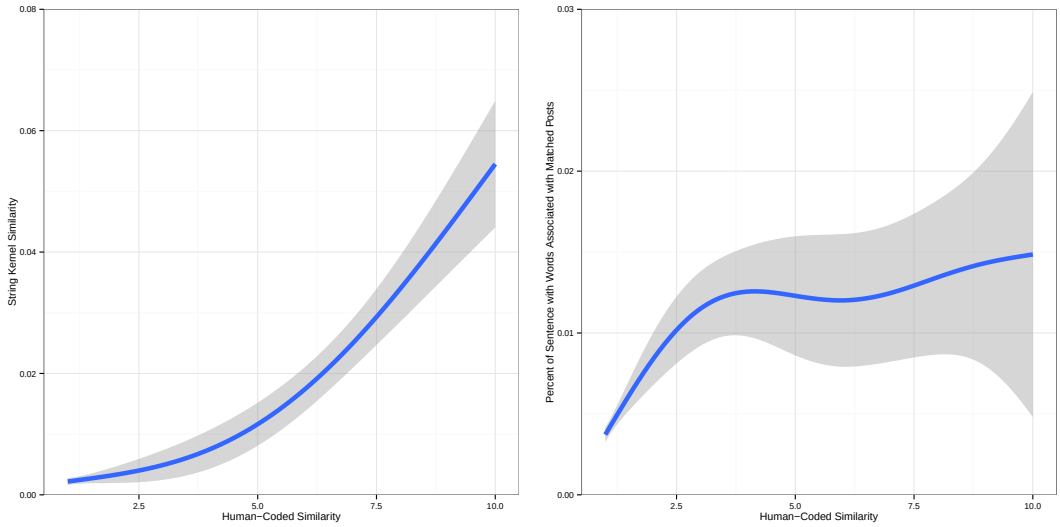


Figure 12: Relationship Between Human Coded Similarity and (1) String Kernel Similarity and (2) Word Overlap

References

Edmond, Chris. 2013. "Information manipulation, coordination, and regime change." *The Review of Economic Studies* p. rdt020.

Fu, King-wa, Chung-hong Chan and Marie Chau. 2013. "Assessing censorship on microblogs in China: Discriminatory keyword analysis and the real-name registration policy." *Internet Computing, IEEE* 17(3):42–50.

Gomez, James. 2000. *Self censorship: Singapore's shame*. Think Centre.

Hassanpour, Navid. 2011. "Media disruption exacerbates revolutionary unrest: Evidence from Mubaraks natural experiment." *APSA 2011 Annual Meeting Paper* .

Hassid, Jonathan. 2010. "Pressing Back: The Struggle for Control over China's Journalists.".

Hassid, Jonathan. 2012. "Safety valve or pressure cooker? Blogs in Chinese political life." *Journal of Communication* 62(2):212–230.

Howard, Philip N and Muzammil M Hussain. 2011. "The role of digital media." *Journal of Democracy* 22(3):35–48.

Huang, Haifeng. 2013. "Propaganda and Signaling." *Available at SSRN 2325101* .

Iacus, Stefano M., Gary King and Giuseppe Porro. 2009. "CEM: Coarsened Exact Matching Software." *Journal of Statistical Software* 30. http://gking.harvard.edu/cem.

Jin, Qiu. 1999. *The culture of power: the Lin Biao incident in the Cultural Revolution*. Stanford University Press.

Kalathil, Shanthi and Taylor C Boas. 2010. *Open networks, closed regimes: The impact of the Internet on authoritarian rule*. Carnegie Endowment.

Kelly, Sanja, Sarah G Cook, Mai Truong and Freedom House. 2012. *Freedom on the Net 2012: A Global Assessment of Internet and Digital Media*. Freedom House.

King, Gary, Jennifer Pan and Margaret E. Roberts. 2013. "How Censorship in China Allows Government Criticism but Silences Collective Expression." *American Political Science Review* 107:1–18. http://j.mp/LdVXqN.

Lee, Francis LF and Angel MY Lin. 2006. "Newspaper editorial discourse and the politics of self-censorship in Hong Kong." *Discourse & Society* 17(3):331–358.

Lessig, Lawrence. 1999. *Code: And other laws of cyberspace*. Basic Books (AZ).

Link, Perry. 2009. "The Anaconda in the Chandelier: Censorship in China Today." *Scholars Under Siege? Academic and Media Freedom in China* .

Lorentzen, Peter. 2010. "Regularizing Rioting: Permitting Protest in an Authoritarian Regime." Working Paper.

Manning, Christopher D, Prabhakar Raghavan, Hinrich Schütze et al. 2008. *Introduction to information retrieval*. Vol. 1 Cambridge university press Cambridge.

Morozov, Evgeny. 2012. *The net delusion: The dark side of Internet freedom*. PublicAffairs.

Qiang, Xiao. 2011. The Rise of Online Public Opinion and Its Political Impact. In *Changing Media, Changing China*, ed. Susan Shirk. New York: Oxford University Press pp. 202–224.

Roberts, Margaret E, Brandon M Stewart and Richard Nielsen. 2015. "Matching Methods for High-Dimensional Data with Applications to Text.".

Stern, Rachel E and Jonathan Hassid. 2012. "Amplifying silence uncertainty and control parables in contemporary China." *Comparative Political Studies* 45(10):1230–1254.

Wacker, Gudrun. 2003. The Internet and censorship in China. In *China and the Internet: Politics of the digital leap forward*, ed. Christopher R Hughes and Gudrun Wacker. Routledge.

Zheng, Yongnian. 2007. *Technological empowerment: The Internet, state, and society in China*. Stanford University Press.