RESEARCH ARTICLE

POLITICAL SCIENCE

Reverse-engineering censorship in China: Randomized experimentation and participant observation

Gary King,[1]* Jennifer Pan,[1] Margaret E. Roberts[2]

Existing research on the extensive Chinese censorship organization uses observational methods with well-known limitations. We conducted the first large-scale experimental study of censorship by creating accounts on numerous social media sites, randomly submitting different texts, and observing from a worldwide network of computers which texts were censored and which were not. We also supplemented interviews with confidential sources by creating our own social media site, contracting with Chinese firms to install the same censoring technologies as existing sites, and—with their software, documentation, and even customer support—reverse-engineering how it all works. Our results offer rigorous support for the recent hypothesis that criticisms of the state, its leaders, and their policies are published, whereas posts about real-world events with collective action potential are censored.

The Chinese government has implemented "the most elaborate system for Internet content control in the world" (*1*), marshaling hundreds of thousands of people to strategically slow the flow of certain types of information among the Chinese people. Yet the sheer size and influence of this organization make it possible to infer via passive observation a great deal about its purpose and procedures, as well as the intentions of the Chinese government. To get around the well-known inferential limitations inherent in observational work, our experiment depended on large-scale random experimentation and participant observation.

We begin with the theoretical context. The largest previous study of the purpose of Chinese censorship (*2*) distinguished between the "state critique" and "collective action potential" theories of censorship and found that, with few exceptions, the first was wrong and the second was right: Criticisms of the government in social media (even vitriolic ones) are not censored, whereas any attempt to physically move people in ways not sanctioned by the government is censored. Even posts that praise the government are censored if they pertain to real-world collective action events (*2*).

In both theories, regime stability is the assumed ultimate goal (*3*–*6*). Scholars had previously thought that the censors pruned the Internet of government criticism and biased the remaining news in favor of the government, thinking that others would be less moved to action on the ground as a result (*7*–*9*). However, even if biasing news positively would in fact reduce the potential for collective action, this state critique theory of censorship misses the value to the central Party organization of the information content provided by open criticism in social media (*10*–*13*). After all, much of the job of leaders in an autocratic system is to keep the people sufficiently mollified that they will not take action that may affect their hold on power. In line with the literature on responsive authoritarianism, the knowledge that a local leader or government bureaucrat is engendering severe criticism—perhaps because of corruption or incompetence—is valuable information (*14*, *15*). That leader can then be replaced with someone more effective at maintaining stability, and the system can then be seen as responsive. This responsiveness would seem likely to have a considerably larger effect on reducing the probability of collective action than merely biasing the news in predictable ways.

The collective action potential hypothesis holds that the Chinese censorship organization first detects a volume burst of social media posts within a specific topic area, and then identifies the real-world event that gives rise to the volume burst (*2*). If the event is classified as having collective action potential, then all posts within the burst are censored, regardless of whether they are critical or supportive of the state and its leaders. Unlike the uncertain process involved in coherently classifying individual posts as to their collective action potential, this procedure is easily implemented with extremely high levels of intercoder reliability. No evidence exists as to whether any such rules were invented and directed by a person or committee in the Chinese government, or whether they merely represent an emergent pattern of this large-scale activity.

Although the largest existing study analyzed more than 11 million social media posts from almost 1400 websites across China (*2*), it and other quantitative studies of censorship (*16*, *17*) were solely observational, with some conclusions necessarily depending on untestable assumptions. For example, the data for these studies were controlled by an earlier stage in which many social media websites use automated review (based on techniques such as keyword matching) to immediately move large numbers of prospective posts into a temporary limbo to receive extra scrutiny before possible publishing (for a guide, see Fig. 1). Whereas the ex post content-filtering decision is conducted largely by hand and takes as long as 24 hours, the ex ante decision of whether posts are slotted for review is automated, instantaneous, and thus cannot be detected by observational methods. This also means that the automated review process could induce selection bias in existing studies of censorship, which could only observe those submissions that were not stopped from publication by automated review. And, of course, observational research generally also risks endogeneity bias, confounding bias, and other problems.

To avoid these potential biases and to study how automated review works, we conducted a large-scale experimental study in which random assignment, controlled by the investigators, substituted for statistical assumptions. We created accounts on numerous social media sites across China; wrote a large number of unique social media posts; randomized the assignment of different types of posts to accounts; and, to evade detection, observed from a network of computers all over the world which types were published or censored. Throughout, we attempted to avoid disturbing the flow of normal discourse by producing social media content on topics similar to those in real social media posts (including the content of those censored, which our methods could access). Although very-small-scale nonrandomized efforts to post on Chinese websites and observe censorship have been informative (*18*), randomized experiments have not before been used in the study of Chinese censorship.

In addition to our randomized experiment, from which we drew causal inferences, we also sought to produce more reliable descriptive knowledge of how the censorship process works. This is important information in its own right; the process is intensely studied and contested in the academic and policy communities. Until now, such information has mostly come from highly confidential interviews with censors or their agents at social media sites or in government. This information is necessarily partial, incomplete, potentially unsafe for research subjects, and otherwise difficult to gather. Participant observation provided us with a new source of information absent from previous studies of censorship. From inside China, we created our own social media website, purchased a URL, rented server space, contracted with one of the most popular software platforms in China used to create these sites, submitted, automatically

[1]Institute for Quantitative Social Science, Harvard University, Cambridge, MA 02138, USA. [2]Department of Political Science, University of California, San Diego, La Jolla, CA 92093, USA.
*Corresponding author. E-mail: king@harvard.edu

reviewed, posted, and censored our own submissions. The website we created was available only to our research team, so as to avoid affecting the object of our study or otherwise interfering with existing Chinese social media discourse. However, we had complete access to the software, documentation, help forums, and extensive consultation with support staff; we were even able to get their recommendations on how to conduct censorship on our own site in compliance with government standards. The "interviews" we conducted in this way were unusually informative because the job of our sources was in fact to answer the questions we posed.

Overall, this work offers three intended contributions. First, by analyzing large numbers of posts at numerous social media sites, we are able to resolve some disagreements in the policy and academic literatures on the subject, such as explanations for the presence of conflicting keyword lists and the absence of a coherent or unified interpretation for the operation of these lists at individual sites. Consistent with this disagreement, we show that the large number of local social media sites in China have considerable flexibility, and choose diverse technical and software options, in implementing censorship. Second, we show that the automated review process affects large numbers of posts on fully two-thirds of Chinese social media sites, but is a largely ineffective step in implementing the government's censorship goals. This is surprising but consistent with the known poor performance of most keyword-based approaches to text classification. Finally, despite automated review's large presence, high potential for generating selection bias in observational studies, and overall ineffectiveness due to keyword matching, we find that the government is still able accomplish its objectives—as summarized by the collective action potential hypothesis—by using very large numbers of human coders to produce post hoc corrections to automated review and to censorship in general.

Our research offers clear support for the collective action potential hypothesis and then offers some important extensions. We find—consistent with the implications of this theory, but untested in prior research—that there is no censorship of posts about collective action events outside mainland China, collective action events occurring solely online, social media posts containing critiques of top leaders, and posts about highly sensitive topics (such as Tibet and Xinjiang) that do not occur during collective action events.

Research designs

We now describe the challenges involved in large-scale experimentation, participant observation,

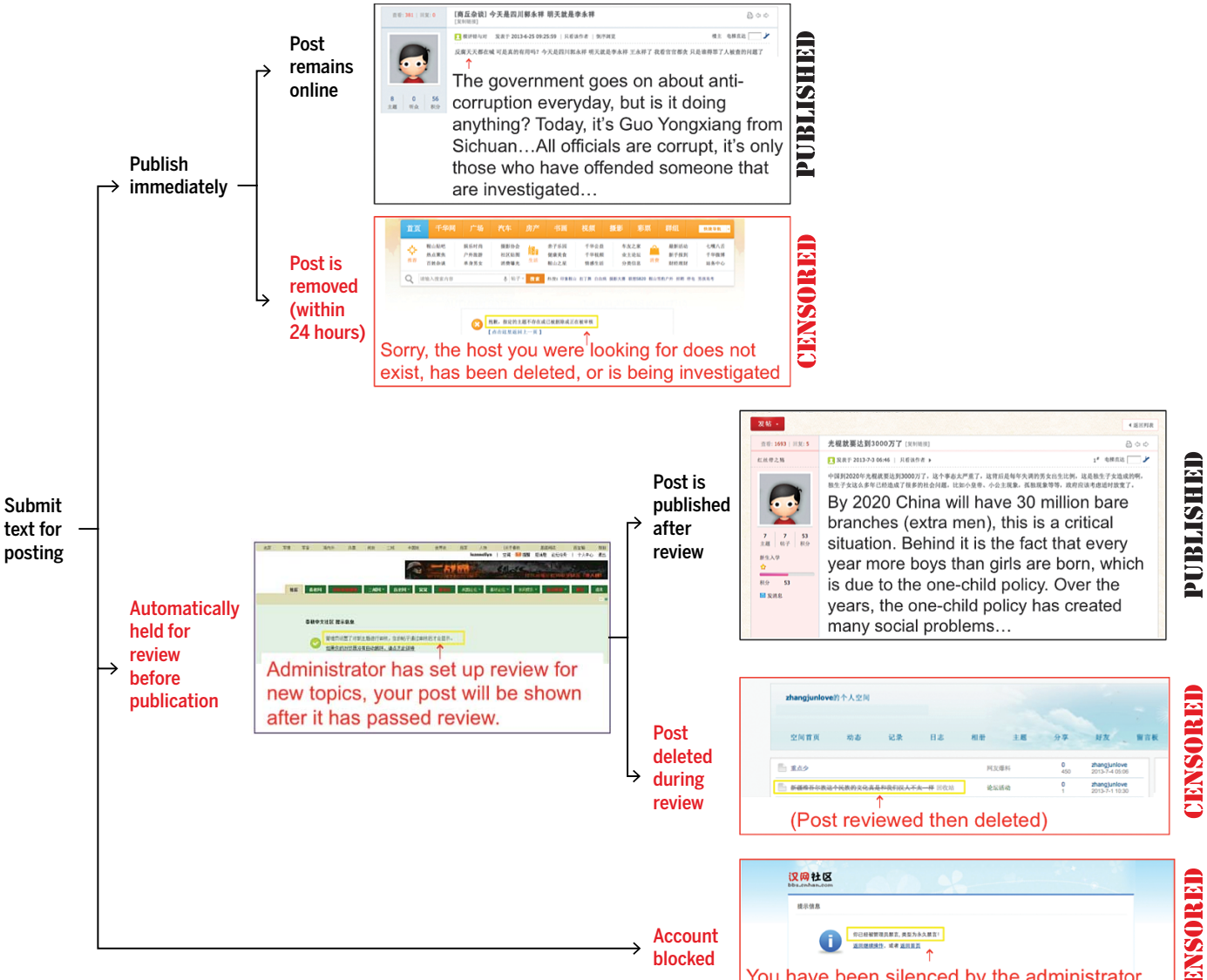


Fig. 1. The Chinese censorship decision tree. The pictures shown are examples of real (and typical) websites, along with our translations. Observational studies are based only on the first three paths through this decision tree; our experimental study includes all five. Full screen shots are in (*19*).

and data collection in a system designed to prevent the free flow of information, especially about the censors [see (*19*) for additional details]. These include avoiding detection, implementing the experiment in many geographically distant places, keeping a large research team safe, and ensuring that we did not disturb or alter the system we were studying. The human subjects aspects of our experimental protocol were preapproved by our university's institutional review board (IRB). For obvious reasons, we are unable to reveal certain details of how we implemented this design, but we do give complete information on the statistical and scientific logic behind our choices (*20*).

We begin with the outcome variable we are studying and then describe our experimental protocols.

Participant observation

Aspects of the process by which censors in the Chinese government and social media companies implement censorship directives have been gleaned over the years in interviews with sources who have first-hand knowledge, including the censors themselves. We have also conducted many such interviews, and each one produces some information, but it is necessarily a partial picture, highly uncertain, and potentially unsafe for the sources and researchers.

Thus, we looked for a way to learn more by changing the incentives of our sources. We did this by creating our own Chinese social media site from inside China, using all the infrastructure, procedures, and rules that existing sites must follow. We purchased a URL, contracted with a company that provides hosting services, and arranged with another company to acquire the software necessary to establish a community discussion forum [a bulletin board system (BBS)]. We downloaded the software and installed it ourselves. This infrastructure gave us complete access to the software and its documentation so that we could fully understand and make use of its functionality. Support employees at these firms were happy to help show us how to censor in such a way as to maintain our website in accordance with their view of government requirements. Thus, instead of trying to convince people to spare some of their time for researchers, we were able to have conversations with employees whose job was to answer questions like those we posed; fortunately, they seemed quite good at this. We then studied and customized the software, submitted posts ourselves, and used the software's mechanisms to censor some of them. We took every step we could (short of letting individuals in China post on the site) to avoid causing any interference to actual social media discourse.

The biggest surprise we found relative to the literature was the huge variety of technical methods by which automated review and human censorship can be conducted. Table 1 summarizes some of these options.

When we installed the software, we found that, by default, it included no automated review or blocking. But webmasters can easily change the option of automatically reviewing specific types of users (those who are moderators, super users, users who have been banned from posting, or those who have been banned from visiting the site), Internet protocol (IP) addresses, new threads, or every response—all of which can be tailored for each of as many forums as desired on each website. Functionality also exists to bulk-delete posts, which can be implemented by date range, user name, user IP, content containing certain keywords, or length of post. On the back end, the webmaster also has flexible search tools to examine content, search by user name, post titles, or post content. What the user sees can also be limited: The search function can be disabled for users, and webmasters have the option of allowing users to see whether posts of theirs are being automatically reviewed (and, if so, which ones).

We found employees of the software application company to be forthcoming when we asked for recommendations as to which technologies have been most useful to their other clients in following government information management guidelines. On the basis of their recommendations as well as user guides, detailed analyses from probing the system, and additional personal interviews (with sources granted anonymity), we deduce that most social media websites that conduct automatic review do so via a version of keyword matching, probably using hand-curated sets of keywords (we reverse-engineer the specific keywords below) (*21*).

We summarize our understanding of the censorship process in Fig. 1. The process begins when one writes and submits a blog or microblog post at a social media website (left). This post is either published immediately (top left node) or held for review before publication (middle left node in red). If the post is published immediately, it may be manually read by a censor within about 24 hours and, depending on the decision, either remains online indefinitely (top box) or is removed from the Internet (second box). As can be seen from the screen shots of actual websites in Fig. 1 [full examples in (*19*)], the decisions of the censors, and the fact that they are making these decisions, are unambiguous.

The censors then read each post in review (usually within a day or two) and either publish the post (third box of Fig. 1) or delete it before publication (fourth box). We are able to identify review when it occurs because typically the user receives a message after post submission that the text has been slotted for review. In the absence of a warning message, the user can tell when a post is put into review because no public URL is associated with the post, and the user's account page will show the status of the post as "under review." Finally, on the basis of the current and previous posts, a submitted post can be censored and the account blocked so that no additional posts may be made (last box of Fig. 1). In this last case, when a user submits a text for

Table 1. Options for content filtering on forum platform.

Automated review options	
Content-based review can be based on:	• Moderator-supplied keywords • Plug-ins for reviewing posts with minimal influence on the user • Plug-ins advertising better keyword-blocking technology • Review specific to post type (e.g., comment or main post) • Review specific to forum topic
User-based review can be based on:	• User IP • Payments by user • Points won by user (e.g., for number of posts, comments) • Previous user posts • Last login
Time-period review and censorship allows:	• Periods of time when all posts are audited • Prevention of posting during certain hours of the day
Workflow for reviewed posts:	• Different censors for different types of postings (e.g., spam versus political content) • Batch deletion of posts • Review interface with search functionality
Account blocking options	• Blocking for specific types of posts (e.g., comment or main post) • Blocking for specific forums • Blocking based on points • Blocking based on user IP • Blocking posting and/or reading

posting, an error message notifying the user of account blocking is encountered. A key point is that the massive data set in (2) corresponds only to the first three boxes, whereas in our experiment we are able to study all five paths down the decision tree.

Experimental protocol

We designed our experimental protocol to make causal inferences without certain modeling assumptions. We first selected 100 social media sites, including 97 of the top blogging sites in the country, representing 87% of blog posts now on the web. We included the top three microblogging (i.e., Twitter-like) sites: Sina Weibo (weibo.com), Tencent Weibo (t.qq.com), and Sohu Weibo (t.sohu.com). The first two of these microblogging sites each have more than 500 million registered users and 50 to 100 million daily active users (22). Together, the 100 sites are geographically spread all over China; 20 are run by the government, 25 are state-owned enterprises, and 55 are private firms. Some cater to national audiences, whereas some only allow posting within a local area. Creating accounts on some of these sites requires the user to be in the country at a specific geographic locale, to have a local e-mail address, or to provide another method of communication for identification. We devised procedures to create two accounts at each of these 100 social media sites.

We kept our design close to aspects of (2). The theory in that paper was not that every social media post with the potential to generate collective action is censored; after all, almost any issue could in principle be used as a hook to generate protest activity. Instead, the theory is that pro- or anti-government posts concerning a collective action event are censored. Collective action events are those "which (a) involve protest or organized crowd formation outside the Internet; (b) relate to individuals who have organized or incited collective action on the ground in the past; or (c) relate to nationalism or nationalist sentiment that have incited protest or collective action in the past" [(2), p. 6].

We conducted three rounds of experiments (18 to 28 April, 24 to 29 June, and 30 June to 4 July 2013) during which social media posts would need to be written in real time about current issues. This presented a logistical challenge. At the beginning of each round, we scoured the news and selected ongoing collective action events and non–collective action events about which there was a volume burst in social media discussion. We chose a ratio of one collective action event to two non–collective action events, because collective action events are more scarce and so that we could average over different non–collective action events. We included non–collective action events only if they were widely discussed topics pertaining to actions taken by the Chinese government, officials, or the Chinese Communist Party (CCP) that were unrelated to events with collective action potential. We also attempted where possible to select events that mentioned specific officials' names and addressed what has been described as especially "sensitive" topics. (We also included several edge cases, described below.) Details of all events appear in (19), but here are the four collective action events we found when our study was conducted, all of which meet the definition but some of which are more incendiary than others:

1. Qui Cuo, a 20-year-old mother, self-immolated to protest China's repressive policies over Tibet. Her funeral drew protesters.

2. Protesters in Panxu, a village in Xiamen Fujian, took to the streets because they claimed officials did not adequately compensate them for requisitioning their collectively owned farmland to build a golf course. Village representatives went to local authorities to demand compensation but were instead detained. Thousands of villagers went to the town hall to demand the release of the village representatives, police moved in to arrest the villagers, and the villagers retaliated by smashing police cars and taking the local Party secretary into custody.

3. On the second anniversary of the 2011 arrest of artist-dissident Ai Weiwei, he released a music album that talked about his imprisonment. Ai Weiwei was arrested in 2011 on charges of tax evasion, but more likely the true reason was either that he called upon his followers to mimic the Arab Spring or that he organized volunteers to collect the names of children who died in the Sichuan earthquake. The release of the album by Ai Weiwei is chosen as an example of collective action under part (b) of the definition, where posts about individuals who have organized or incited collective action on the ground in the past are censored.

4. An altercation between protesting Uyghurs (a minority ethnic group) and local police in Lekeqin township of Shanshan county in Turpan, Xinjiang, led to the deaths of 24 people, including 16 Uyghurs. Police and many official news reports of the event termed it an act of Uyghur terrorism, but rumors circulated in social media that the protest was precipitated by forced housing demolition.

For each event, we had a group of native Chinese speakers write some posts supportive and others critical of the government. These posts were based on social media posts that had already appeared online, including posts that were censored as well as those that remained online. [We used the technology of (2) to obtain access to the censored posts.] In other words, we obtained posts that were immediately published after submission, including those that remained online and those that were removed (top two boxes of Fig. 1). We provided our writers with background on the event, the definition of what we meant by pro- and anti-government for each topic (19), and examples of real posts from Chinese social media similar to those we needed written. So that we could minimize any experimenter effect, we checked each text ourselves by hand and attempted throughout to ensure that the posts we submitted were similar in language, sentiment, and content to those already found in (or written and censored in) Chinese social media.

From a statistical point of view, we ensured balance by blocking (23) on (that is, randomly sampling only within each cell of the cross-classification of) three variables: First, our posts included the same keywords in both the treatment and control conditions. Second, we controlled for individual writing style by blocking on author in our experimental design. That is, posts in each set of four experimental conditions (defined by our two variables: pro- or anti-government, and with or without collective action potential) were written by the same set of research assistants. Finally, we constrained all posts to be between 100 and 200 characters in length. In addition, we also ensured that no two posts submitted were exactly identical to each other or to any we found in social media. All posts were submitted between 8 a.m. and 8 p.m. China time, either from the United States or from the appropriate place within China, depending on what was feasible because of the technology used at each social media site (24).

We were interested in testing the causal effect of both pro- versus anti-government content and collective action versus non–collective action content, leading by cross-classification to four logical treatment categories. To make the most efficient use of each individual account, we submitted two posts to each. But it makes little sense for one account (representing a single person) to write both pro- and anti-government posts regarding the same event. Thus, we submitted posts about two different events to each account; some of these posts were pro-government collective action and anti-government non–collective action, and others were anti-government collective action and pro-government non–collective action. In this way, every account contributed to the causal effect estimate of each hypothesis. We also ensured our ability to make causal inferences without extra modeling assumptions by randomizing (i) the choice between these two pairs, (ii) the order within each pair, and (iii) the specific collective action and policy events we wrote about in each submission. Missingness could occur when websites were down, if an account we created expired, or if an account had been blocked because of prior posts. Largely because of the design, any missingness is almost exactly independent of our two treatment variables; empirically that proved to be the case.

Each of the 100 different social media websites in our study offers a different way of expressing oneself online. When possible, we submitted posts on the home page we created for each account. For discussion forums, we started a new thread with the content of the post in the most popular subforum. On sites where users are not permitted to create new threads, we submitted posts as a reply to an existing thread relevant to the topic. In all cases, we wrote our posts so as not to stand out from the stream of existing information; we followed all social media, website, and cultural norms. In total, we wrote 1200 posts by hand, every one unique, and none referring to each other (25).

After submitting a post, we observed whether it was caught by the process of automated review; if in automated review, whether and when it was eventually published; and if not caught by the automated review process, whether it was eventually censored after the fact or remained on the web. When a post appeared on the web, we recorded the URL and verified censorship from computers inside and outside of China. We recorded the outcome in terms of censorship, which corresponds to the branches of the decision tree in Fig. 1.

Throughout, our goal was that anyone looking at our submissions would have no idea that this was part of an academic research project, and that the posts were typical of what anyone might otherwise find online and would not in any way disrupt or change the social media ecosystem we were studying. We also needed to ensure that our checking of published posts for censorship was not obtrusive. So far as we are aware, no one outside of our research team and confidants were aware of this experiment before now, and no one on the web indicated any suspicion about or undue attention toward any of our posts.

Results

We find that in aggregate, automated review affects a remarkably large portion of the social media landscape in China. In total, 66 of the 100 sites in our sample (automatically) review at least some social media submissions, and 40% of all of our individual social media submissions from our 100 sites (and 52% of submissions from sites that review at least sometimes) are put into review. Of those submissions that go into review, 63% never appear on the web.

These figures indicate that automated review affects a large component of intended speech in China and so deserves systematic attention from researchers. This is especially so because of conflicting conclusions and lack of a unified interpretation in the academic and policy literatures about which keywords provoke action by the government, how automated review works, and what impact this process ultimately has on the content of speech that is blocked and that which can be consumed by the Chinese people (*26*, *27*). We offer a possible resolution to these issues here.

Censorship

Using our broader sample, unaffected by selection during the automated review process, and with our experimental randomization, we began by testing the collective action potential hypothesis. On the basis of a difference in means between the treatment and control groups, the black dots in the left panel of Fig. 2 summarize the point estimate for the causal effects on censorship of submitting posts about four separate collective action events. The vertical lines are 95% confidence intervals (as with all our figures). The effects are substantial, ranging from about 20 to 40 percentage point differences (denoted on the vertical axis) solely due to writing about an ongoing collective action event as compared to an ongoing non–collective action event.

We also examined some of the other decision paths in Fig. 1. To do this, we estimated the "causal mediation effect" (*28*, *29*) of submitting posts about collective action events (versus non–collective action events) on censorship, and found that almost none of this effect is mediated through automated review: The overall effect is a trivial 0.003 probability, with a 95% confidence interval of (−0.007, 0.016) (*19*). The (non)effect for each of the four collective action events we studied is displayed in the right panel of Fig. 2, and each is similarly approximately zero, with a small confidence interval. Review, which appears to be fully automated, is thus applied in a manner independent of other relevant variables, and, like most keyword-only methods of automated text analysis, it does not work well when applied to large numbers of documents. From this result, it even appears that the censors largely ignore it, or at least do not get much information from it (see below).

In parallel to the large causal effect for collective action, Fig. 3 reports tests of the state critique hypothesis for each of our four collective action events and eight (non–collective action) policy events. The black dots summarize point estimates of the causal effect of submitting posts in favor of the government versus opposed to the government about each event. As can be seen, the dots are all very close to the horizontal dashed line, drawn at zero effect, with six dots above and six below, and all but one of the confidence intervals crossing the zero line. Note especially that there is no hint of more censorship of anti-government posts when they involve topics that might be viewed as more sensitive or which specifically mention the names of Chinese leaders [see (*19*) for contextual details]. This finding runs counter to anecdotal evidence that rumors and names of leaders unrelated to collective action lead to censorship.

Automated review

The overall results in favor of the collective action potential hypothesis and against the state critique hypothesis thus appear unambiguous. The automated review process has a nearly undetectable effect on evidence about that hypothesis, because the human censors correct errors after the keyword-matching techniques are applied in automated review (although even incorrect keyword filtering slows down communications on many subjects). We now go back up the decision tree of Fig. 1 to study the automated review process more directly.

We first noticed that not all websites have automated review turned on, and that the method of censorship varies enormously by website [this is also true for account blocking (*19*)]. This

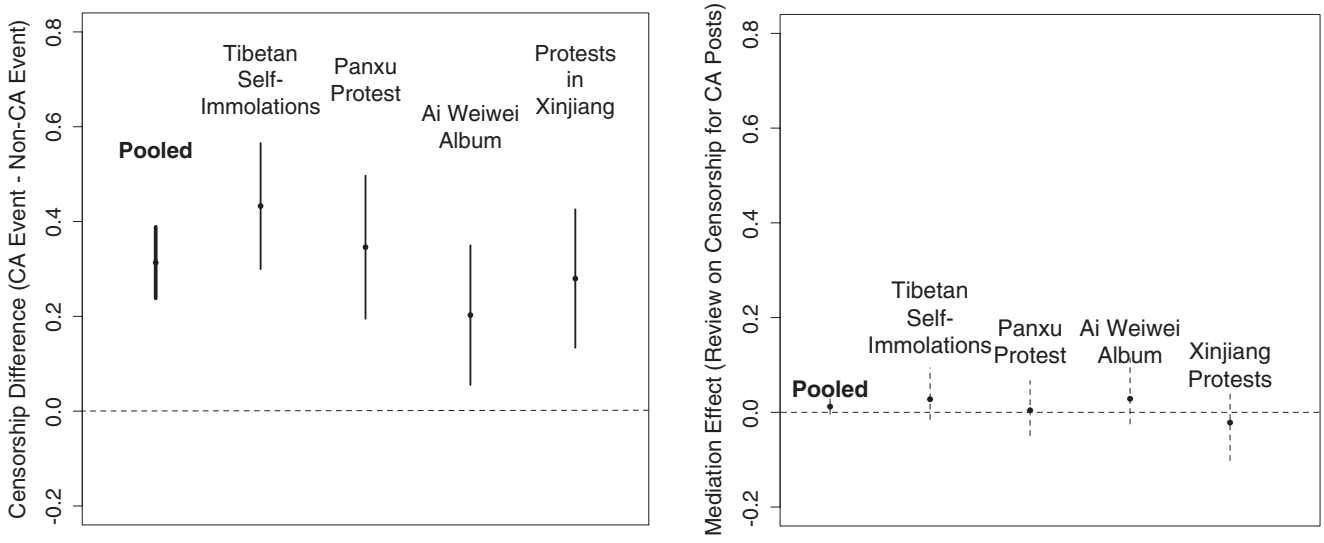


Fig. 2. The causal effect on censorship of posts with collective action potential (left panel) and the mediation effect of review (right panel). Collective action events are more highly censored than non–collective action events within the same time period. However, censorship of collective action events is not mediated through automated review.

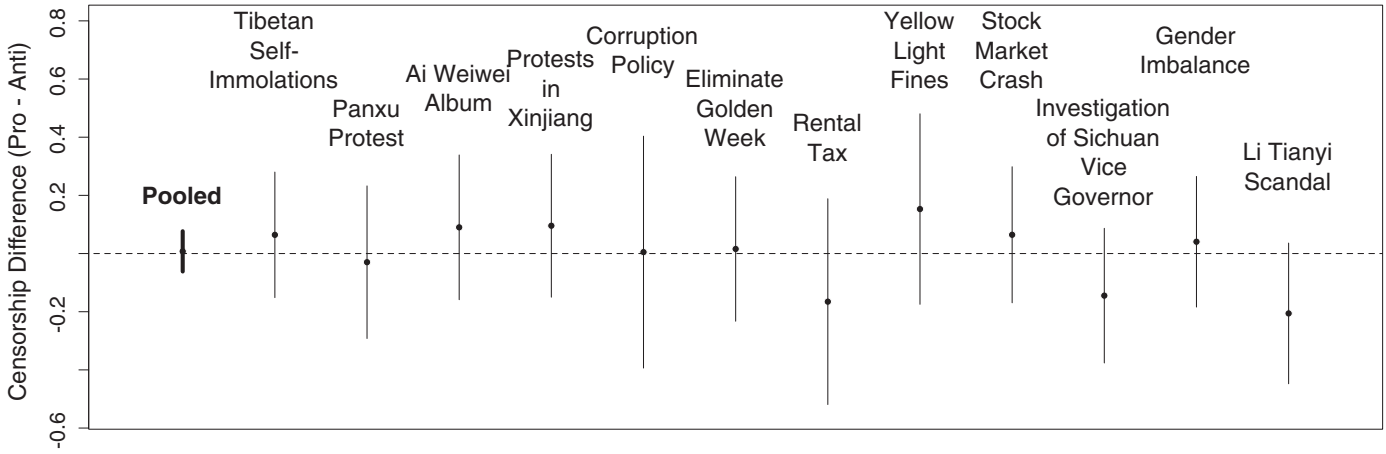


Fig. 3. The causal effect on censorship of posts for or against the government. Posts that support the government are not more or less likely to be censored than posts that oppose the government, within the same topic.

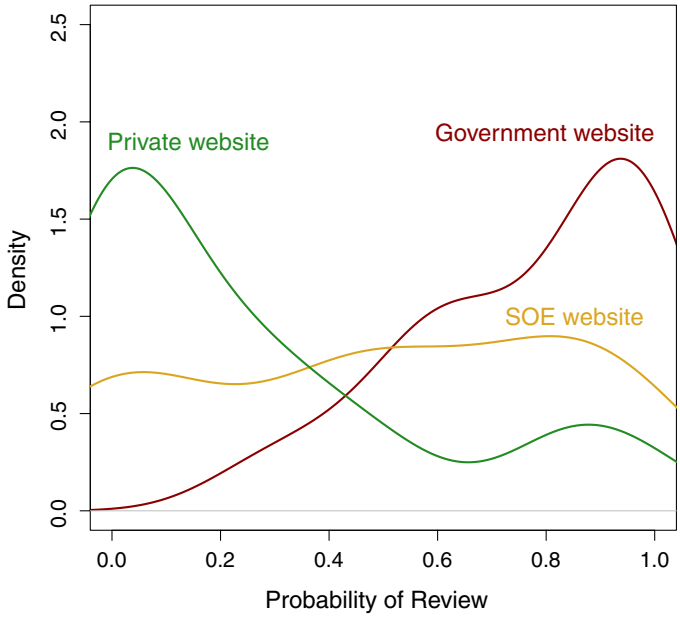
is consistent with what we learned from creating our own social media site, where the software platform not only allows the option of whether to review, but also offers a large variety of choices of the criteria by which to review. Indeed, there exists considerable diversity in the technologies used by different social media sites for automated review (*17*). It is this diversity in technology across sites, then, that appears to account for why different researchers typically find different patterns when looking at different sites or at specific issues. This also accounts for why researchers have been unable to offer unified interpretations of their observations that are consistent with reasonable assumptions about the goals of the Chinese leadership. Only by looking at the whole process does the simplicity of the Chinese government's goals become clear.

Why would the government, in the course of providing top-down authoritarian control, allow for a free choice from a large number of censorship methods? To answer this question, we collected detailed information about all software platforms and plug-ins available for purchase or license by social media sites to control information. From this study, we conclude that the government is (perhaps intentionally) promoting innovation and competition in the technologies of censorship. Such decentralization of policy implementation as a technique to promote innovation is common in China (*30*–*33*).

On the basis of interviews with those involved in the process, we also found a great deal of uncertainty over the exact censorship requirements and the precise rules under which the government would interfere with the operation of social media sites, especially for smaller sites with limited government connections. This uncertainty is in part a result of encouraging innovation, but it may also in some situations be a means of control as well; it is easier to keep people away from a fuzzy line than from a clearly drawn one.

Our systematic empirical study began by investigating which social media websites use any automated review process. Figure 4 presents a density estimate (a continuous version of a histogram) of the distribution of the proportion of posts reviewed for three types of sites, depending on ownership. As can be seen, it is government sites that have the highest probability of a post being put into review, followed by the state-owned enterprises, followed last by privately owned sites (which tend to have the largest user bases).

Fig. 4. Histogram (density estimate) of the proportion of posts reviewed by site. The graph shows that government-controlled social media sites catch many more posts by automated review than do privately owned sites; social media sites controlled by state-owned enterprises (SOEs) are in the middle.



Why would government sites be more likely to delay publication until after automated review, whereas private sites publish first and make censorship decisions later? So far as we can tell from qualitative evidence, the reason is that the penalty for letting offending posts through differs between government and private sites. A government worker who fails to stem collective action could lose his or her job immediately; in contrast, a worker in a private site who makes the same mistake cannot usually be directly fired by the government. Indeed, government workers have a historical legacy of giving priority to following orders and not making mistakes, even if it is considerably more inefficient to do so (*34*). Private social media sites, on the other hand, have incentives to publish as much as they can, so as to attract more users. A private site can, of course, be taken down entirely, but that kind of "nuclear option" is used less often than more generalized pressure on the leadership of these private sites.

What are these largely government sites reviewing? In a manner directly parallel to Figs. 2 and 3 for the ultimate variable of censorship, we analyzed the effects on automated review of collective action and pro- and anti-government posts. Figure 5 gives results for the effect of collective action on review; they include four positive estimated effects, but two are small and three have zero inside their confidence intervals. If the goal of the censors is to capture collective action events, the automated algorithm is performing marginally at best, although this is quite

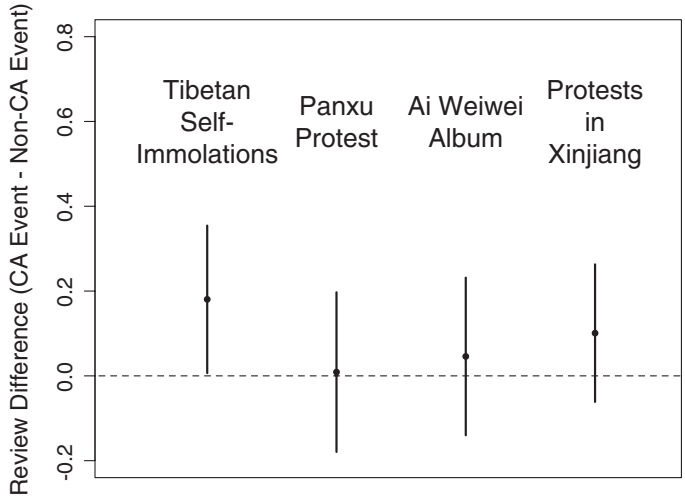


Fig. 5. Causal effect on review of collective action potential events. Collective action events are overall slightly more likely to be reviewed than non–collective action events.

common for keyword algorithms, which tend to work well for specific examples for which they can be designed but often have low rates of sensitivity and specificity when used for large numbers of documents.

Also interesting is the causal effect of pro- versus anti-government posts in Fig. 6. These are all small, and most of the confidence intervals cross zero. If there exists a nonzero relationship here, it is that submissions in favor of the government are reviewed more often than those against the government. Indeed, 9 of 12 point estimates are above zero, and two even have their entire confidence interval above zero. This presents a mystery: Government social media sites are slightly more likely to delay publication of submissions that favor the government, its leaders, or their policies. Private sites do not use automated review much at all. Why is this? We found that the answer again is the highly inexact keyword algorithms used to conduct the automated review.

To understand this better, we reverse-engineered the Chinese keyword algorithms in order to discover the keywords that distinguish submissions reviewed from those not reviewed. Because the number of unique words written overwhelms the number of published posts, we could not find these keywords uniquely. However, we could identify words highly associated with review using a "term frequency, inverse document frequency" algorithm (*35*, *36*). That is, we took the frequency of each word within the review posts and divided this number by the number of nonreviewed documents in which that same word appeared. Thus, for every word we obtained a measure of its frequency in review posts, relative to posts that were not reviewed. Words with high values on these measures are likely to be used within the automated review process.

Table 2 gives the top keywords (and key phrases) that we estimate were used to select posts we wrote into automated review. We can see that the words associated with review could plausibly detect collective action and relate to the government and its actions, but are also just as likely to appear in pro-government posts as in anti-government posts. For example, more pro- than anti-government posts are reviewed in the Corruption Policy topic in Fig. 6. This appears to be because the reviewed pro-government posts used the word corruption (腐败) more frequently than did anti-government posts. However, corruption was used in the context of praising how the new policy would strengthen anti-corruption (反腐败) efforts. Not only is automated review conducted by only a subset of websites and largely ineffective at detecting posts related to collective action events, it also can backfire by delaying the publication of pro-government material.

It turned out that we could provide an independent test of the veracity of these keywords. In the context of setting up our own website, we unearthed a list of keywords for review that a software provider offered to its clients running social media websites. The list is dated April 2013, and all of the keywords we found related to events taking place prior to April 2013 were on this list; the exceptions were from events that occurred after April 2013.

It thus appears that the workers in government-controlled websites are so risk-averse that they have marshaled a highly error-prone methodology to try to protect themselves. They apparently know not to take this automated review methodology very seriously; whether it is used or not, the manual process of reading individual posts must still be used widely, as our results show that automated review does not affect the causal effect of collective action events on censorship decisions.

Edge cases

We now attempt to define the outer boundaries of the theory of collective action potential by examining cases close to but outside the theory (where no effect is anticipated), as well as one extreme case inside the theory: criticism of the top leaders.

Internet-only and external-only collective action

The first case is an event in which collective action took place, but only on the Internet. At the end of May 2013, the principal of Hainan Wanning City No. 2 Elementary School was being investigated for taking six elementary school girls to a hotel. Ye Haiyan, a women's rights advocate, went to the elementary school and protested with a sign in her hand that read "Principal: Get a hotel room with me, let the elementary students go! Contact telephone: 12338 (Ye Haiyan)." Ye's protest went viral and her sign became an online meme, where individuals would take and share photos of themselves, holding a sign saying the same thing with their own phone numbers or often with China's 911 equivalent (110) as the contact phone number (*37*).

The second event occurred on 1 July 2013, which was the 16th anniversary of the handover of sovereignty of Hong Kong from Britain to China. Every year on this day, thousands take to the streets of Hong Kong in protest, but typically with little or no such protest on the mainland. In 2013, between 30,000 people (according to the police) and 430,000 people (according to the organizers) took to the streets to call for true democracy and Chief Executive C. Y. Leung's resignation (*38*).

Neither of these "edge case" examples meet the definition of collective action events given above, but they are obviously close. We ran our experimental design for these events too (Fig. 7, left panel). In both cases, the overall causal effect is near zero, with confidence intervals that overlap zero. There is a hint of a possibly positive effect only for posts reviewed about Hong Kong protests, but in the context of the natural variability of Figs. 2 and 3, this effect is not obviously different from zero.

Corruption and wrongdoing among senior leaders

Next, we consider the effects of writing about corruption and wrongdoing among senior leaders in the government, Party, and military on censorship. Nothing in the theory of collective action potential supports this effect, but because corruption so directly implicates leaders who could control censoring, considerable suspicion exists in the literature that posts about corruption are censored (*16*, *18*, *26*). We can even point to the odd result that posts supporting the government's efforts to deal with corruption are more censored than posts opposed to the government's efforts to deal with corruption (see Fig. 6).

We selected three corruption-related topics for the analysis. The first relates to a new corruption policy that imposes criminal charges against bribes exceeding 10,000 Chinese yuan. The second topic relates to the investigation of Guo Yongxiang, a member of the Sichuan Province Central Committee and a Vice Governor of Sichuan, for serious breaches in discipline. The final topic relates to the naming of Li Tianyi, the son of the well-known People's Liberation Army performer Li Shuangjiang, for participating in a gang rape. The Li Tianyi case led to speculations of corruption that Li's father's ties to the People's Liberation Army would allow Li to avoid punishment commensurate with his crimes. The results for an analysis of the three corruption events

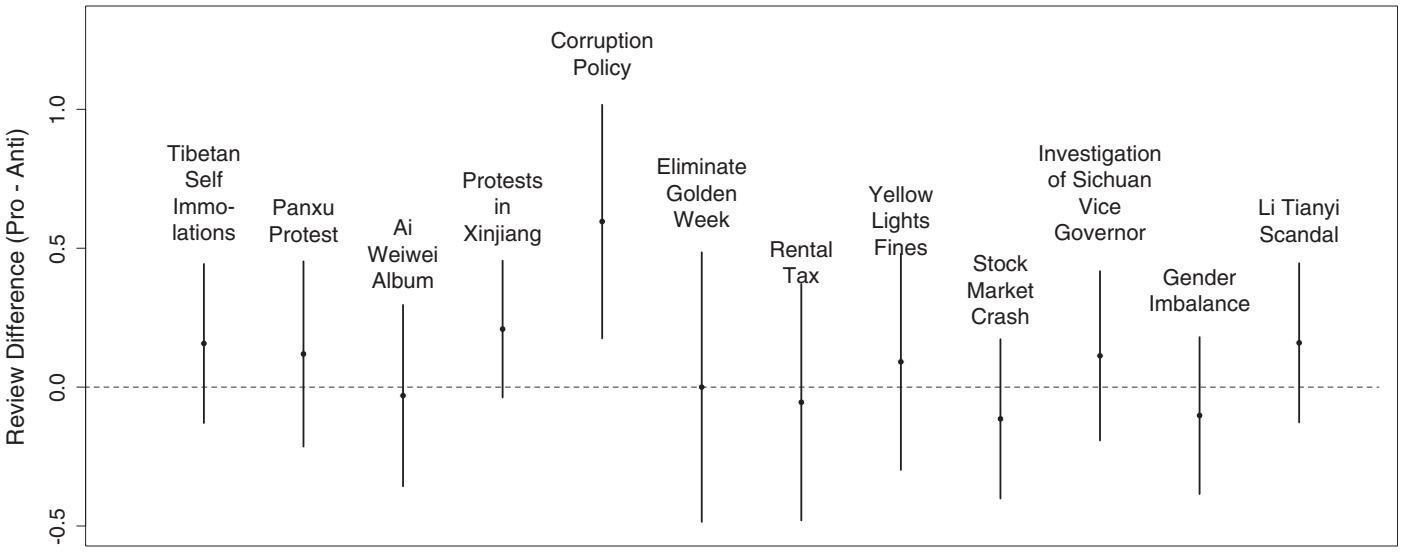


Fig. 6. Causal effect on review of posts for or against the government. Automated review picks up slightly more posts that are for the government as compared to posts that are against the government.

Table 2. Top keywords distinguishing posts held versus not held for review. Words within this list match keyword lists provided by the software provider.

Chinese	English
群众	masses
政府	government
事件	incident
恐怖	terror
新疆	Xinjiang
中国	China
上街	go on the streets
李天一	Li Tianyi
法律	law
达赖	Dalai Lama
游行	demonstration
香港	Hong Kong
行贿	to bribe
腐败	corruption

(Fig. 7, right panel) clearly show no effect, thus again supporting the theory of collective action potential. Similarly supportive is the fact that posts in these topics name specific Chinese government and CCP leaders, both at central and local levels of government (*19*).

Top leaders and highly sensitive issues

Finally, we used observational methods to study the question of the censorship of discussions about top Chinese leaders, arguments for deep political reform, and discussion of highly sensitive or salient issues.

To study more directly whether Chinese censors allow direct criticism of top leaders, we began by finding a social media volume burst about Chinese President Xi Jinping that (i) by our specific definitions does not have collective action potential, (ii) includes posts that cover meaningful and important topics, and (iii) is about a topic that could generate highly critical posts about the leader. We found the following volume burst that met these conditions.

On 28 December 2013, President Xi Jinping visited a Feng Qing Steamed Bun Shop in Beijing (Feng Qing is a chain restaurant) and ate steamed buns "just like the rest of us." He waited in line, he paid 21 CNY for steamed pork and onion buns along with a side of stir-fried liver, and he brought his own tray to a table. Xi's visit unleashed a storm of traditional media coverage and a large volume burst on social media. Although Xi's visit to the bun shop sounds like an innocuous event, online discussions related the visit to important and high-profile issues such as Xi's China Dream, corruption of government officials, rising real estate prices, and the plight of China's elderly and impoverished, as well as propaganda, censorship, the absence of elections, and multiparty competition. However, this event is not connected to any ongoing collective action events.

During this volume burst, we collected 82,280 social media posts related to this event before any posts were censored, and then checked each one from a network of computers around the world that were eventually censored. Finally, we applied the Hopkins-King algorithm (*39*) (using a training set of 592 hand-coded posts) to determine the proportion of censored posts that were critical versus supportive, and applied the Bayesian algorithm derived in (*2*) to invert this. We found, consistent with the collective action potential hypothesis, that posts critical of President Xi were censored just about as much as those that were supportive. Among posts that were critical of Xi and his actions, 18% were censored (95% confidence interval, 13 to 22%). Among the posts supportive of Xi, 14% were censored (95% confidence interval, 8 to 22%). [The proportion of posts censored among posts that simply described the event was 21% (95% confidence interval, 18 to 24%).]

The supplementary materials (*19*) include the text of examples of uncensored posts that are highly critical of President Xi and that use this event to discuss important issues. These posts involve many vivid personal attacks on Xi and his policies. In our experience, these posts are not surprising or unusual.

Next, we looked for uncensored discussion of deep political reform. In August 2013, three commentaries were published in *People's Daily* condemning constitutionalism, describing constitutionalism as incompatible with socialism and doomed to fail in China. These commentaries sparked a social media volume burst with intensive online discussions about whether China should adopt American-style constitutionalism and multiparty competition. In the days after these commentaries, we collected a random sample of 9850 blog posts related to political reform. Although this sample includes posts that toe the party line and criticize constitutionalism, there are also many uncensored posts advocating for the adoption of multiparty competition, describing reform as the only way to empower the Chinese people and to rein in corruption. We include several examples in (*19*).

Finally, we sought and identified social media volume bursts related to three highly salient and politically sensitive issues about real-world events that did not have collective action potential. These are discussions related to Tibet, Uyghurs, and Ai Weiwei.

First is the case of a volume burst in Tibet: In early August 2013, a post by a woman who claimed to have spurned her true love in order to marry a man who lived within view of Lhasa's Potala Palace went viral. As expected, censorship of posts in this burst was low at 12%.

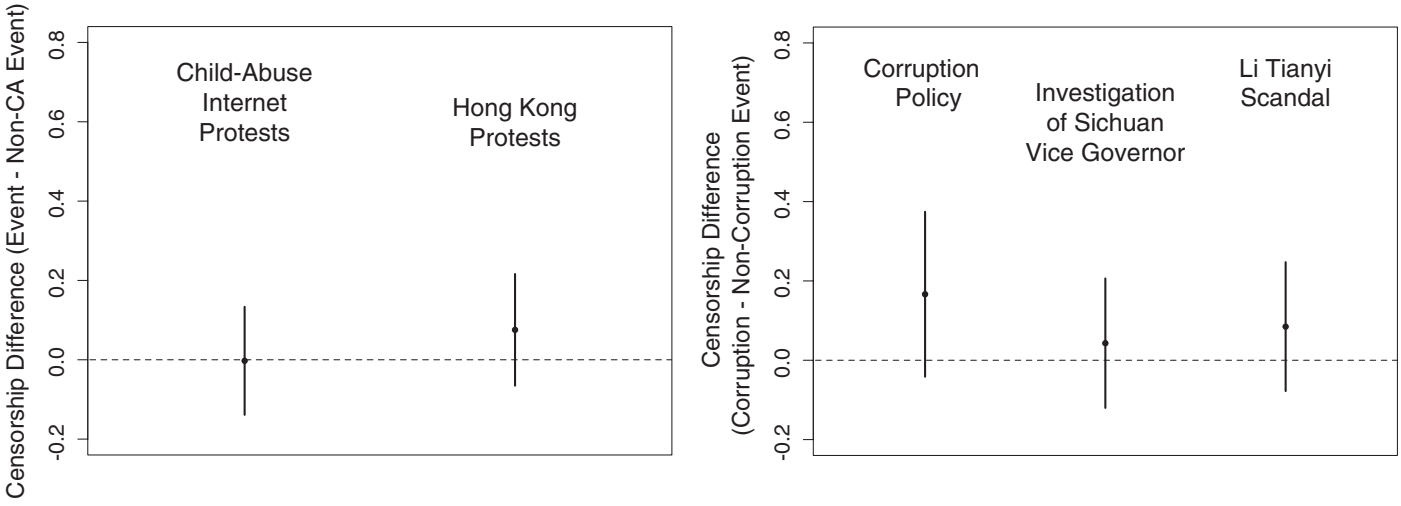


Fig. 7. Testing edge cases for the causal effect of collective action potential (left panel) and of posts about corruption (right panel).

Second is a volume burst related to Xinjiang and Uyghurs, which occurred in March 2013 when a post poking fun at a government entity with an exceptionally long, 54-character name (新疆维吾尔自治区乌鲁木齐国家高新技术产业开发区社会管理综合治理委员会学校及周边治安综合治理工作领导小组办公室) went viral. The government entity is located in Xinjiang, and the name can be roughly translated as the "Public Security and Management Office of the Working Small Group of the Holistic Social Management Committee's School and Surrounding Areas of Xinjiang Uyghur Autonomous Region Urumqi's Chinese High Tech Development Zone." This post was the butt of jokes and satire related to Chinese government bureaucracy but was completely unrelated to any ongoing collective action event. As expected, censorship of this volume burst was low, only 10%.

Finally, we identified a volume burst related to artists—including Ai Weiwei along with Matisse, Picasso, Andy Warhol, and others—and their cats. Censorship of this burst was also low, at 6%. Our definition of collective action potential includes real-world events related to those who have catalyzed or organized collective action in the past. This volume burst relates to Ai Weiwei but falls outside the definition because the burst is not related to a real-world event, nor is it solely related to Ai Weiwei.

Concluding Remarks

We offer the first large-scale randomized experimental analysis of censorship in China, along with participant observation of how censorship is conducted. We use these designs to conduct a rigorous test of the theory of collective action potential, and to further uncover and resolve academic conflicts about crucial aspects of the Chinese censorship program. With them, we are able to subject to empirical estimation what had previously been left to statistical assumption. We are also able to study the large program whereby enormous numbers of social media submissions are put into limbo before being considered for possible publication or censorship. Whereas censorship is a "publish first, censor later" process, automated review involves a "review first, maybe publish later" process.

Our flexible research designs also enabled us to study edge cases, just beyond the reigning theory of collective action potential, so that we can define the boundaries of where it applies. This includes the effects of highly salient and sensitive topics about events without collective action potential; posts about corruption; posts that name Chinese leaders specifically; and collective action events that are solely on the Internet or about collective action on the ground outside the Chinese mainland—none of which are predicted by the theory of collective action potential to be censored more than others, and which our data clearly show are not censored more than other non–collective action topics. We also show that academic controversies over confusing interpretations of which keywords are being censored in automated review are resolved once we realize that the Chinese government is surprisingly flexible concerning what methods and technology each social media site can use, even while imposing uniformity of results by requiring post hoc censoring by human coders.

Future researchers should consider comparing these results on censorship in social media with censorship in traditional media and other ways the Chinese government impedes the free flow of information.

REFERENCES AND NOTES

1. Freedom House, "Freedom of the press, 2012"; www.freedomhouse.org.
2. G. King, J. Pan, M. E. Roberts, How censorship in China allows government criticism but silences collective expression. *Am. Pol. Sci. Rev.* **107**, 326–343 (2013). doi: 10.1017/S0003055413000014
3. S. L. Shirk, *China: Fragile Superpower: How China's Internal Politics Could Derail Its Peaceful Rise* (Oxford Univ. Press, New York, 2007).
4. S. L. Shirk, *Changing Media, Changing China* (Oxford Univ. Press, New York, 2011).
5. M. K. Whyte, *Myth of the Social Volcano: Perceptions of Inequality and Distributive Injustice in Contemporary China* (Stanford Univ. Press, Stanford, CA, 2010).
6. L. Zhang, A. Nathan, P. Link, O. Schell, *The Tiananmen Papers* (Public Affairs, New York, 2002).
7. A. Esarey, Q. Xiao, Political expression in the Chinese blogosphere: Below the radar. *Asian Surv.* **48**, 752–772 (2008). doi: 10.1525/AS.2008.48.5.752
8. R. MacKinnon, *Consent of the Networked: The Worldwide Struggle For Internet Freedom* (Basic Books, New York, 2012).
9. P. Marolt, Grassroots agency in a civil sphere? Rethinking internet control in China. In *Online Society in China: Creating, Celebrating, and Instrumentalising the Online Carnival*, D. Herold, P. Marolt, Eds. (Routledge, New York, 2011), pp. 53–68.
10. M. Dimitrov, The resilient authoritarians. *Curr. Hist.* **107**, 24–29 (2008).
11. P. L. Lorentzen, Regularizing rioting: Permitting public protest in an authoritarian regime. *Q. J. Pol. Sci.* **8**, 127–158 (2013). doi: 10.1561/100.00012051
12. P. L. Lorentzen, China's strategic censorship. *Am. J. Pol. Sci.* **58**, 402–414 (2014). doi: 10.1111/ajps.12065
13. X. Chen, *Social Protest and Contentious Authoritarianism in China* (Cambridge Univ. Press, New York, 2012).
14. E. Malesky, P. Schuler, Nodding or needling: Analyzing delegate responsiveness in an authoritarian parliament. *Am. Pol. Sci. Rev.* **104**, 482–502 (2010). doi: 10.1017/S0003055410000250
15. G. Distelhorst, Y. Hou, Ingroup bias in official behavior: A national field experiment in China. *Q. J. Pol. Sci.* **9**, 203–230 (2014). doi: 10.1561/100.00013110
16. D. Bamman, B. O'Connor, N. Smith, Censorship and deletion practices in Chinese social media. *First Monday* **17** (no. 3) (March 2012). doi: 10.5210/fm.v17i3.3943
17. T. Zhu, D. Phipps, A. Pridgen, J. Crandall, D. Wallach, The velocity of censorship: High-fidelity detection of microblog post deletions. In *22nd USENIX Security Symposium* (Washington, DC, 14 to 16 August 2013); www.usenix.org/conference/usenixsecurity13/technical-sessions/paper/zhu.
18. R. MacKinnon, China's censorship 2.0: How companies censor bloggers. *First Monday* **14** (no. 2) (February 2009). doi: 10.5210/fm.v14i2.2378
19. See supplementary materials on *Science* Online.
20. We also added our own ethics rules, not required by the IRB, which dictate that we avoid, wherever possible, influencing or disturbing the system we are studying (*19*). The similarity to the Prime Directive in *Star Trek* notwithstanding, this seems like the appropriate stance for scientists attempting to understand the world, as distinct from advocates trying to change it, and in any event is more likely to yield reliable inferences.
21. In the process of setting up the site, they recommended that we hire two or three censors for every 50,000 users. That enables us to back out an estimate of the total number of censors hired within firms at between 50,000 and 75,000, not

counting censors within government, 50 Cent Party members, or the Internet police.
22. See (*40*, *41*) for numbers of registered users, which are substantial even if we account for automated sites created by marketing firms (*42*).
23. K. Imai, G. King, E. Stuart, Misunderstandings between experimentalists and observationalists about causal inference. *J. R. Stat. Soc. Ser. A* **171**, 481–502 (2008). doi: 10.1111/j.1467-985X.2007.00527.x
24. All posts were made to mainland China accounts. Some were submitted from outside China, when feasible, and many from within China. Recent work has noted that overseas accounts are subject to less stringent censorship regulations than mainland accounts (*43*). This issue does not affect our work because all accounts created and used were mainland China accounts. Users could control account location when creating the account by specifying a location in China, by entering a local mobile number, or by creating the account from a local IP address. We used all of these methods.
25. For each of our three rounds, we wrote 200 posts on non–collective action events (split equally between pro- and anti-government) and 200 posts on collective action events or edge cases (again split equally between pro- and anti-government). Thus, 600 posts submitted relate to non–collective action events, and 600 relate to collective action events or edge cases. We have in total four collective events and two edge cases, and so 400 posts focused on collective action events and 200 on edge cases.
26. J. Crandall *et al*., Chat program censorship and surveillance in China: Tracking TOM-Skype and Sina UC. *First Monday* **18** (no. 7) (July 2013). doi: 10.5210/fm.v18i7.4628
27. J. Fallows, The connection has been reset. *Atlantic* (March 2008); www.theatlantic.com/magazine/archive/2008/03/-the-connection-has-been-reset/306650.
28. K. Imai, L. Keele, D. Tingley, T. Yamamoto, Unpacking the black box of causality: Learning about causal mechanisms from experimental and observational studies. *Am. Pol. Sci. Rev.* **105**, 765–789 (2011). doi: 10.1017/S0003055411000414
29. J. Pearl, Direct and indirect effects. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence* (Morgan Kaufmann, San Francisco, 2001), pp. 411–420; https://dslpitt.org/uai/papers/01/p411-pearl.pdf.
30. O. Blanchard, A. Shleifer, "Federalism with and without political centralization: China versus Russia" (National Bureau of Economic Research, Cambridge, MA, 2000); www.nber.org/papers/w7616.
31. S. Heilmann, E. Perry, *Mao's Invisible Hand: The Political Foundations of Adaptive Governance in China* (Harvard University Asia Center, Cambridge, MA, 2011).
32. Y. Qian, G. Roland, Federalism and the soft budget constraint. *Am. Econ. Rev.* **88**, 1143–1162 (1998).
33. Y. Qian, B. R. Weingast, Federalism as a commitment to perserving market incentives. *J. Econ. Perspect.* **11**, 83–92 (1997). doi: 10.1257/jep.11.4.83
34. G. Egorov, K. Sonin, Dictators and their viziers: Endogenizing the loyalty-competence trade-off. *J. Eur. Econ. Assoc.* **9**, 903–930 (2011). doi: 10.1111/j.1542-4774.2011.01033.x
35. G. Salton, *Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer* (Addison-Wesley, Reading, MA, 1988).
36. D. Kelleher, S. Luz, Automatic hypertext keyphrase detection. In *Proceedings of the 19th International Joint Conference on Artificial Intelligence* (Erlbaum, Hillsdale, NJ, 2005), pp. 1608–1609.
37. For examples, see (*44*).
38. For news coverage of the protests, see (*45*–*48*).
39. D. Hopkins, G. King, A method of automated nonparametric content analysis for social science. *Am. J. Pol. Sci.* **54**, 229–247 (2010). doi: 10.1111/j.1540-5907.2009.00428.x
40. K. Hong, China's Twitter-like Sina Weibo service now has over 50 million active users per day. *The Next Web*, 13 August 2013; http://tnw.co/1fdNFPS.
41. S. Millward, Tencent Weibo, the 'other weibo' that nobody cares about, reaches 540 million users. *Tech in Asia*, 22 January 2013. http://bit.ly/1byjSNW.
42. K. W. Fu, M. Chau, Reality check for the Chinese microblog space: A random sampling approach. *PLOS ONE* **8**, e58356 (2013). doi: 10.1371/journal.pone.0058356; pmid: 23520502
43. J. Q. Ng, Weibo keyword un-blocking is not a victory against censorship. *Tea Leaf Nation*, 21 June 2013; http://bit.ly/1kfqNBC.
44. P. Barefoot, "Principal, get a room with me, spare the schoolchildren!" *China Smack*, 31 May 2013; http://j.mp/19yuv7E.
45. Al-Jazeera, Democracy push as Hong Kong marks handover. 1 July 2013; http://j.mp/145Jvpp.
46. S. Lee, K. Wong, Hong Kong protests to underscore Leung's record-low appeal. *Bloomberg BusinessWeek*, 28 June 2013; http://j.mp/13r3v7v.
47. J. Ngo, July 1 protest is Hong Kong's taste of democracy. *South China Morning Post*, 30 June 2013; http://j.mp/15PcwBt.
48. C. Yung, Annual Hong Kong protest to focus ire on leader. *Wall Street Journal*, 28 June 2013; http://j.mp/13FJB3w.
49. G. King, J. Pan, M. E. Roberts, Replication Data for: Reverse Engineering Chinese Censorship: Randomized Experimentation and Participant Observation (2014). doi: 10.7910/DVN/26212

ACKNOWLEDGMENTS

For helpful advice, we thank P. Bol, S. Chestnut, P. Gries, Y. Herrera, H. Huang, I. Johnston, S. Shirk, D. Tingley, and participants in a panel at the American Political Science Association meeting, 31 August 2013, and at the Midwest Political Science Association meeting, 3 April 2014. For expert research assistance over many months, we are tremendously appreciative of the efforts and insights of F. Chen, W. Cheng, A. Jiang, A. Jin, F. Meng, C. Li, H. Liu, J. Sun, H. Waight, A. Xiang, L.-S. Xu, M. Yu, and a large number of others whom we shall leave anonymous. We thank Crimson Hexagon Inc. for help with data. See (*49*) for replication data and information.

SUPPLEMENTARY MATERIALS

www.sciencemag.org/content/345/6199/1251722/suppl/DC1
Materials and Methods
Figs. S1 to S7
References (*50*, *51*)

3 February 2014; accepted 2 July 2014
10.1126/science.1251722

www.sciencemag.org/content/345/6199/1251722/suppl/DC1

Supplementary Materials for

Reverse-engineering censorship in China: Randomized experimentation and participant observation

Gary King,* Jennifer Pan, Margaret E. Roberts

*Corresponding author. E-mail: king@harvard.edu

Published 22 August 2014, *Science* **345**, 1251722 (2014)
DOI: 10.1126/science.1251722

This PDF file includes:

Contents

1 Introduction

This paper includes supplementary materials for this paper:

> Gary King; Jennifer Pan; and Margaret E. Roberts. 2014. "Reverse Engineering Chinese Censorship: Randomized Experimentation and Participant Observation."

Detailed replication information and data for the paper can be found in:

> Gary King; Jennifer Pan; Margaret E. Roberts, 2014, "Replication data for: Reverse Engineering Chinese Censorship: Randomized Experimentation and Participant Observation", http://dx.doi.org/10.7910/DVN/26212, UNF:5:K/LGmB0vjskGYBobxbT+8g== IQSS Dataverse Network [Distributor] V3 [Version]

2 Topic Details

Here we offer details about the collective action and non-collective action events we found and used in the paper. Also included are the two edge case events. We list the events used within each of the three rounds of our experiment, by round.

Round 1

- CA event 1, Tibetan Self-immolation: Pro-government posts attribute the tragedy of her death on the Dalai Lama who is instigating these tragedies, Anti-government posts attribube her death to government policies.
- CA event 2, Protest in Panxu village over illegal land seizure: Pro-government posts say that this sort of protest and violence is wrong and that the villagers are greedy and want money. Anti-government posts say the local officials are unfair to the villagers.
- Non-CA event 1 Corruption Policy: new policy that bribes over 10,000 Chinese yuan will be subject to criminal investigation and penalities. Pro-government posts support this policy because it will reduce corruption. Anti-government posts believe this policy is punishing those who give bribes but the real fault lies with officials

who accept bribes and not those who are forced by the system to give bribes in order to get things done.

- Non-CA event 2, Eliminate Golden Week: people were calling for removal of the 10 day holiday that occurs during China's National Day. Pro-government posts support the 10 day holiday, saying that it stimulates domestic consumption, tourism revenues, stimulates economic development, and allows everyone to relax to promote social harmony. Anti-government posts call for removal of the policy because millions of people traveling at the same time is unsafe and unsanitary and the government should heed the call of the many poeple who are calling for the government to abolish the Golden Week holiday.

- Non-CA event 3, Rental tax: several cities in China are piloting taxes for renting housing (charging taxes on their rental income), which stimulated a lot of discussion and debate. Pro-government posts support the rental tax because it is income that should be taxed, just as income from salaries and wages are taxed. Anti-government posts criticize the tax saying it will increase already high rental taxes as landlords will push the tax onto renters.

- Non-CA event 4, Yellow Light fines: China promulgated new traffic regualtions, which generated debate, especially the part that running yellow lights will incur punishment and fines. This debate prompted the authorities to say that punishment will be in the form of education, not fines or harsher penalties. Pro-government supports the new policy because it will improve transportation safety, and says that education not punishment is what's needed. Anti-government rejects and criticizes the authorities for not upholiding the spirit of the law (i.e., education is not punishment).

Round 2

- CA event 1, Dissident Ai Weiwei releases a new album called Divine Comedy: Pro-government criticizes Ai Weiwei for releasing the album. Anti-government supports Ai Weiwei's actions and the album.

- Non-CA event 1, Shanghai Stock Market crash: Steep decline in the Shanghai stock market (the largest single day decline in the past four years). Pro- government says the government has done everything it can to regulate financial markets and this crash is the work of speculators and hackers. Anti-government posts say the stock market crashed and caused hardship to ordinary investors because of bad government interventions, policies, and actions.

- Non-CA event 2 (Corruption), Investigation of Sichuan Vice-Governor Guo Yongxiang: Guo is being investigated for serious breaches of discipline (i.e., corruption). Guo was a member of the Sichuan Province Standing Committee and a Vice Governor. Pro- government says the investigation is good because it will cut down on corruption. Anti-government says all officials are corrupt and Guo is being investigated for other political reasons.

- Edge case 1, Online Protest of Child Abuse. Pro-government posts we wrote criticize Ye Haiyan and this form of protest as unproductive and harmful to social order. Anti-government posts support Ye and criticize a corrupt educational system.

Round 3

- CA event 1, Protests in Xinjing. Pro-government posts calls this an act of terrorism against the Chinese people. Anti-government posts say that this event may be due to forced housing demolition instead of terrorism.

- Non-CA event 1 (Corruption), Li Tianyi Scandal: Li Tianyi is the son of a famous People's Liberation Army performer, Li Shuangjiang. The Beijing police department announced that Li Tianyi and four other young men gang raped a young women on Februrary 17, 2013, and that investigation of Li has been completed. Pro-government posts say the government did a good job arresting Li, even though his father is well connected. Anti-government posts say the government is not doing enough, and asks why the other four participants have not been named.

- Non-CA event 2, Gender Imbalance: new report released by the National Statistics bureau says that by 2020, China will have 30 million "bare branches" (extra men). Pro-government says that is the results of backwardness and preference for boys in rural China. Anti-government says that this is the result of the China's one-child policy.

- Edge case 1, Hong Kong protest. Pro-government criticizes these protests are trouble-making and disruption to social harmony. Anti-government says the protests are a means of expression for better government and democracy.

3 Account Blocking

In addition to automated review, and content filtering by censorship, some entire accounts are sometimes blocked, which is another form of information control. We did not design our experiment to study blocking, but we are able to glean some important information about it anyway. Under our experimental design, each social media account we set up ultimately had the same number of collective action related posts. However, blocking can occur at any time, and at different times during our experimental protocol, each account had submitted different numbers of collective action related posts. In addition, censorship of collective action posts was not perfect and so we can also leverage these differences as well. Figure 1 gives the basic relationship among sites that use blocking as a tool. It shows that once the percent censored on an account (see the horizontal axis) hits a rate of at least 60-80%, the probability that that account will be blocked (vertical axis) more than doubles.

We also study whether censorship acts as a mediator between collective action posts and blocked accounts. Using the same methods again, we find an average mediation effect of 0.17 with a 95% confidence interval of (0.09,0.25). This means that censorship alone, independent of content and the collective action content of posts, is what alerts the internet service provider to accounts with collective action content, making them more likely to block the offending account from posting further. Blocking thus appears to be a relatively automated process that is calculated from the number of posts that were censored from

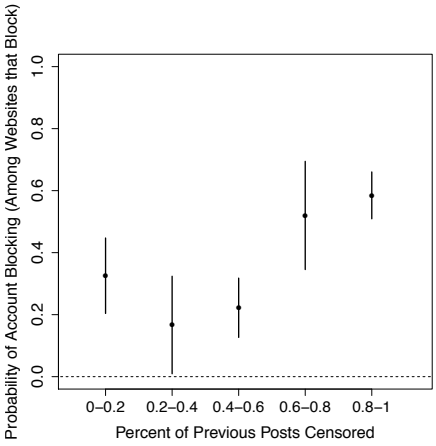


Figure 1: How Blocking is a Function of Prior Censorship

previous attempted posts. It does not seem to be the subject of separate analysis or human judgment in many cases.

4 Examples of the Text of Social Media Posts

We give some posts for the volume burst in two categories: critical posts that are not censored and posts that illustrate that important topics are covered.

Below we include the original chinese text, as well as an English translation. Note, however, that many of these posts are difficult to translate because they use slang, homonyms, and various types of clever wordplays. So we choose to include a literal translation, as well as an explanation for the intended meaning. (A side point is that we translate the food 包子 as bun instead of dumpling, since both are dough products with pork based filling but the bun that Xi ate is steamed, while dumplings are boiled.)

4.1 Examples of Uncensored critical posts about Xi Jinping

> 三国时郭嘉预测孙策必死于刺客之手，朋友问我旁边跟习大大合影的人如果左手 抽刀插入——习大大在福建浙江上海执政的时候从来没有“亲民”，为何到了北 京就变得“亲民”了呢？个人认为仅仅是作秀而已

Literal translation: During the Three Kingdoms period, Guo Jia predicted Sun Ce would die at the hands of an assassin, a friend asked me what would have happened if one

of those people taking pictures with Xi at the dumpling shop had pulled out a knife and stabbed him — when Big Xi was governing in Fujian, Zhejiang, and Shanghai he never got close to the people, why is it that now he's in Beijing he's getting close to the people? In my opinion, he's just putting on a show

Explanation: The second part of this passage accuses Xi of faking his care toward the people, specifically "getting close to the people" means getting to know ordinary citizens and showing them you care.

谁要吧把习大大弄来草包一次，肯定比庆丰包子强

Literal translation: who can make big Xi into a grass bun, it'll definitely be better than Feng Qing buns

Explanation: to make into a grass bun means to make a fool of someone. Big Xi is northern slang for Uncle Xi, a very colloquial way of refering to Xi Jinping. The passage means that to make a fool of Xi Jinping would surely be better than Feng Qing buns.

我觉得关注明星的事情没有什么大不了的，总比关心习总吃什么包子穿 什么衣服彭妈用什么包然后鼓吹个什么光明盛世国货雄起之类的恶心论调好得 多，在那些人眼里他们也不过是消费领导人罢了

Literal translation: I think there's nothing wrong with paying attention to celebrities, and it's infinitely better than paying attention to what kind of bun president Xi is eating, what he is wearing, what handbag Mother Peng (Xi's wife) is using and what bright shiny new product she is promoting, and other sorts of sickening behavior, that shows they (Xi and his wife Peng) are just leaders of consumerism

Explanation: pretty much what the translation says, but the wording, e.g., Mother Peng is so casual that it's clearly making fun of Xi and his wife.

4.2 Examples Illustrating Important Topics

尊敬的习大人，您一顿包子21元，那一天三顿饭光吃包子就算50元好了。 一个月30天，单算全吃包子就要1500元。加上老婆孩子全吃包子总共4500元。 请问上海一个月5000元的工资能干什么？全家都吃包子都不够！

Literal translation: Respected Xi, you ate a meal of buns for 21 CNY, if you ate buns for all the meals, then that would be 50 CNY a day. There are 30 days in a month, if you only ate buns that would be 1500 CNY. If you add in a wife and a child, that's 4500 CNY a month. My question to you: what can you get with Shanghai's monthly wage of 5000 CNY? It's not even enough for a family just eating buns!

> 习总，能不能给无家可归的人吃个包子？“这无家可归的人能吃到包子 吗？？？是什么让她沦落街头？？天朝百姓的包子梦能实现吗？？维稳领导会 不会让她吃锤子、刀子？？？

Literal translation: President Xi, can you give some buns to the homeless? Do the homeless get to eat any buns?? How do these people end up on the streets?? Can the people achieve their bun dreams?? Will leaders who want to maintain social stability force these people (the homeless) to eat hammers and knives???

Explanation: bun dream is a play on Xi's China Dream initiative. Eating hammers and knives refers to the homeless getting beaten up implying China's leaders only care about maintaining stability not improving people's lives.

> 习近平先生，上回你吃包子，我评论了几句，把我的号给封了。这事，我知道不是你直接干的，但是，我想知道你对此类事件的态度。这算是 侵犯公民言论自由权利的行为吗？如果你的回答让我不满意，对不起，我不会 投你的票。习近平先生，告诉你一个秘密，如有选票的话，我真投你的票，原 因嘛，你毕竟让吃喝玩乐的收敛了一些。但是我选举你的意愿没有得到 你的尊重，因为我连选票是什么样子都不晓得，自然我就不高兴了，所以我不 想选举你了 。看到你讲尊重台湾人民，我又有了选举你的意愿，那前提 是你尊重我一回，给一张选票。

Literal translation: Mr. Xi Jinping, last time you ate steamed buns, I wrote a few sentences, and my social media account got shut down. I know you're not directly responsible for that, but I want to know what you think of it (my account getting shut down). Do you regard that as a violation of citizens' rights to freedom of expression? If your answer

doesn't satisfy me, sorry, but I won't vote for you. Mr. Xi Jinping, I want to tell you a secret, if China had elections, I would vote for you, this is because you're kind of reining all that eating, drinking and entertainment. But you don't respect my desire to elect you because I have no idea what a ballot even looks like, so of course I'm not happy so I wouldn't vote for you.

Explanation: Xi has said that Chinese people should have freedom of expression; eating, drinking, and entertainment refers to the corruption of officials.

> 习总昨天去庆丰包子铺吃包子后，各地一把手会不会掀起去街边小吃店 吃饭的高潮泥? 天津书记吃起了 "狗不理" 包子，河北书记吃起了驴肉火烧，陕 西书记吃起了肉夹馍，山东书记吃起了煎饼卷大葱，山西书记吃起了刀削 面。。。2014年小吃年

Literal translation: After President Xi ate Feng Qing steamed buns yesterday, all of the local top leaders are now eating street foods. Tianjin's party secretary ate "goubuli" buns, Hebei party secretary ate donkey fire (a local dish), Shaanxi party secretary age Rou jia mou, Shandong party secretary ate onion pancake rolls, Shanxi Party Secretary ate sliced noodles...2014 is the year of local delicacies.

Explanation: China's local leaders all follow the top leader, even when it's something as ridiculous as eating steamed buns and street food.

> 中国不是奴隶社会，但绝对是奴才社会

Literal translation: China is not a slave society, but it is definitely a society where everyone sucks up to leaders (lackey society).

Explanation: this is a tweet quoting the news that the manager of the Feng Qing shop has preserved the table where Xi sat.

> 包子就是包子，热度高了还是容易露馅

Literal translation: A bun is a bun, if it's cooked too long, the filling will come out

Explanation: this is in response to a post that says the biggest lie occurs when the liar believes what he is saying, which was in turn a post responding to news of Xi Jinping saying China has tried Constitutionalism, parliamentarism, multiparty presidential system,

and none are feasible for China. Basically someone is calling Xi a liar, and this person uses the bun analogy to make the point that eventually (when the bun is cooked too long) the truth (the filling) will be revealed.

4.3 Examples Illustrating Discussions of Deep Political Reform

> 想群众成为真正的主人，只有终结集权专政，建立人手一票、多党竞争 的民主制度

To truly give power to the masses means ending the centralized dictatorship, and adopting multi-party democracy where everyone has a vote.

> 别提反腐了，甚至现在中国的民主、法治、人权都不要再提了，已经在 二十多年前彻底死在“履带”之下了！所谓公审，就是演戏，骗老百姓呢罢了。 在民主、法治、人权相对健全的国家，治理腐败绝对不是靠国家，有效治理腐 败靠的是多党竞争，如果是姓共的监管姓共的，这不是天大的笑话吗！？ 醒来 吧国民！

Don't talk any more about anti-corruption, don't even talk about China's current democracy, rule of law, and human rights; those things were completely destroyed over twenty years ago under the treads of tanks! Our so-called court of law is merely for show, to deceive ordinary citizens. In a country with real democracy, rule of law, and human rights, corruption is not fought by the state; effective anti-corruption efforts rely on multi-party competition; the Party (CCP) providing oversight on the Party is an absolute joke. Wake up Chinese citizens!

> 宪政是多党竞争更是多党监督，在朝可施展抱负在野负监督之责。在 民选国家，公民个人和任何政党，有为才有位有为就有位，没有官位还是公 民，失去执政地位不是失去政治地位和价值。对于一个正常的政党，宪政只会 令其保持生机，绝无令其亡的道理。所谓亡党亡国之忧，罪孽自知罢了。

Constitutionalism is multi-party competition and more importantly multi-party supervision, where parties in office work toward their goals, and parties out of office provide

oversight. In a country with popular elections, citizens and parties with promise gain power, and when out of office, (former) politicians become ordinary citizens, losing political office does not lead one to lose political rights. For any normal political party, losing office doesn't lead to its demise. Those who imply constitutionalism will lead to the demise of the Party and the country are guilty of ignorance.

5 Mediation Analysis

To estimate the influence of automated review in the causal path between collective action and censorship, we follow Imai et al. (2011) in conducting a mediation analysis. We are interested in how the treatment (T), collective action, is mediated through mediator (M), automated review, on an outcome (Y), censorship. Let $Y_i(t)$ denote the potential outcome Y for individual i under treatment T. The causal effect within the potential outcomes framework of the treatment on the outcome is $Y_i(1) - Y_i(0)$, or censorship when the post discusses a collective action event, minus censorship had the post discussed a policy event instead. Let $M_i(t)$ define the potential value of automated review under treatment T for individual i (1 if held for review and 0 if not). The mediation effect of automated review on censorship is: $Y_i(t, M_i(1)) - Y_i(t, M_i(0))$ for treatment value $T = t$. That is, we are interested in the difference between (a) censorship when the post discusses a collective action event and (b) censorship when the post was about a collctive action event, but its review status was as if it were not a collective action event. The idea, then, of mediated causal effects is to control the mediator and see what happens to the outcome variable.

To estimate the average mediation effect, we fit two models, an outcome model and a mediator model. The outcome model is $Y_i = \beta_0 + \beta_1 T_i + \beta_2 M_i + \beta_3 T_i M_i$, and the mediator model is $M_i = \theta_0 + \theta_1 T_i$. We allow for interaction between the mediator and the treatment because it is plausible that the mediation effect is stronger for collective action posts. In a simple linear structual equation model, the average causal mediation effect is $\theta_1(\beta_2 + \beta_3)$.[1]

In order to accurately estimate the mediation effect, the assumption of sequential ig-

[1] We complicate this estimator further, following Imai et al. (2011), to allow for a binary outcome and mediator.

norability must hold. First, the treatment must be ignorable conditional on pretreatment covariates. Because collective action posts are randomized in our experiment, this assumption holds. Second, the mediator must be ignorable conditional on the treatment. We can not think of a covariate that influences both review and censorship, conditional on treatment. Therefore, we believe that this second condition holds.

Following the above procedures, we find and report in the paper that review does not mediate the relationship between collective action posts and censorship. This further consolidates our opinion that review is an ineffective technique in finding what the Chinese government is after — collective action.

6 More Information on Research Ethics

The introduction to Section 2 of the paper discusses IRB approval and the fact that we imposed rules on ourselves that go far beyond those required by the IRB, which only protect research subjects according to the specifics of US law. The main feature of our additional rules was that we attempted to avoid influencing the system we were studying. Adhering to this rule has several purposes. First, it is standard practice in social science to attempt to avoid Hawthorne effects and any other type of investigator-induced change in the system that may confound the intended treatment effect of the experiment. Second, violating this rule would likely increase the probability that the experiment would be interrupted by some of the many social media companies where we submitted posts, or many parts of the Chinese government; this is especially crucial since most large scale social experiments fail for because of "unexpected" political interventions (King et al., 2007). The safety of our large research team was also a continuous concern and the subject of a host of other procedures we followed that we cannot detail here (legitimate researchers are welcome to contact us for more information).

We verified our adherence to the rule, to some degree, by studying whether and how anyone else on Chinese social media responded to our randomly assigned posts. As it turned out, we didn't find a single case in which one of our posts was commented on by another netizen that indicated that they had a hint that the post was part of an experiment.

We also studied the pattern of posts, censoring, and commenting and found it to be close in a variety of ways to that which we obtained by pure observation.

For our participant observation, we were extremely careful to only ask questions that would normally be asked in China by participants in social media in China. Our questions to customer service professionals included no deception. We actually did set up a social media site, we paid for everything we accessed that had a cost, and we did want to know the answers to the questions we asked. Customer service answered questions for us that were similar or identical to those people in Chinese social media companies ask them regularly. We asked no questions about the people answering our questions or about any other human subject. We followed extensive procedures to avoid standing out as different in any way from the thousands of others making inquiries and doing what we did for other purposes. The underlying motivation of almost anyone (including us) posing questions to customer service was not raised, nor was it relevant to the workers. We kept no records of any name or online ID of anyone who provided us information. We believe it would be exceedingly difficult to find out who we spoke to, but even if someone did, what we asked and what they told should be nearly indistinguishable from numerous other interactions that occur in China all the time.

7 Website Screen Shots

The following are full screen shots of websites that censor, review, and block accounts from posting.

Figure 2: Censorship

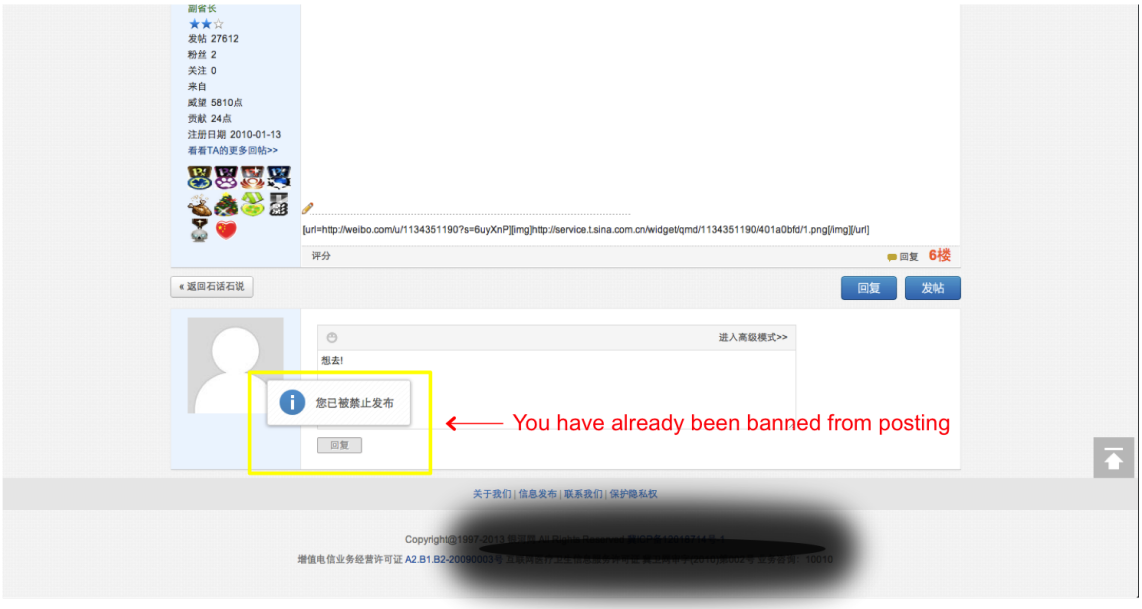
Figure 3: Review

Figure 4: Review

Figure 5: Review

Figure 6: Account Blocking

Figure 7: Account Blocking



References and Notes

1. Freedom House, “Freedom of the press, 2012”; www.freedomhouse.org.

2. G. King, J. Pan, M. E. Roberts, How censorship in China allows government criticism but silences collective expression. *Am. Pol. Sci. Rev.* **107**, 326–343 (2013). doi:10.1017/S0003055413000014

3. S. L. Shirk, *China: Fragile Superpower: How China’s Internal Politics Could Derail Its Peaceful Rise* (Oxford Univ. Press, New York, 2007).

4. S. L. Shirk, *Changing Media, Changing China* (Oxford Univ. Press, New York, 2011).

5. M. K. Whyte, *Myth of the Social Volcano: Perceptions of Inequality and Distributive Injustice in Contemporary China* (Stanford Univ. Press, Stanford, CA, 2010).

6. L. Zhang, A. Nathan, P. Link, O. Schell, *The Tiananmen Papers* (Public Affairs, New York, 2002).

7. A. Esarey, Q. Xiao, Political expression in the Chinese blogosphere: Below the radar. *Asian Surv.* **48**, 752–772 (2008). doi:10.1525/AS.2008.48.5.752

8. R. MacKinnon, *Consent of the Networked: The Worldwide Struggle For Internet Freedom* (Basic Books, New York, 2012).

9. P. Marolt, Grassroots agency in a civil sphere? Rethinking internet control in China. In *Online Society in China: Creating, Celebrating, and Instrumentalising the Online Carnival*, D. Herold, P. Marolt, Eds. (Routledge, New York, 2011), pp. 53–68.

10. M. Dimitrov, The resilient authoritarians. *Curr. Hist.* **107**, 24–29 (2008).

11. P. L. Lorentzen, Regularizing rioting: Permitting public protest in an authoritarian regime. *Q. J. Pol. Sci.* **8**, 127–158 (2013). doi:10.1561/100.00012051

12. P. L. Lorentzen, China’s strategic censorship. *Am. J. Pol. Sci.* **58**, 402–414 (2014). doi:10.1111/ajps.12065

13. X. Chen, *Social Protest and Contentious Authoritarianism in China* (Cambridge Univ. Press, New York, 2012).

14. E. Malesky, P. Schuler, Nodding or needling: Analyzing delegate responsiveness in an authoritarian parliament. *Am. Pol. Sci. Rev.* **104**, 482–502 (2010). doi:10.1017/S0003055410000250

15. G. Distelhorst, Y. Hou, Ingroup bias in official behavior: A national field experiment in China. *Q. J. Pol. Sci.* **9**, 203–230 (2014). doi:10.1561/100.00013110

16. D. Bamman, B. O'Connor, N. Smith, Censorship and deletion practices in Chinese social media. *First Monday* **17** (no. 3) (March 2012). doi:10.5210/fm.v17i3.3943

17. T. Zhu, D. Phipps, A. Pridgen, J. Crandall, D. Wallach, The velocity of censorship: High-fidelity detection of microblog post deletions. In *22nd USENIX Security Symposium* (Washington, DC, 14 to 16 August 2013; www.usenix.org/conference/usenixsecurity13/technical-sessions/paper/zhu).

18. R. MacKinnon, China's censorship 2.0: How companies censor bloggers. *First Monday* **14** (no. 2) (February 2009). doi:10.5210/fm.v14i2.2378

19. See supplementary materials on *Science* Online.

20. We also added our own ethics rules, not required by the IRB, which dictate that we avoid, wherever possible, influencing or disturbing the system we are studying (*19*). The similarity to the Prime Directive in *Star Trek* notwithstanding, this seems like the appropriate stance for scientists attempting to understand the world, as distinct from advocates trying to change it, and in any event is more likely to yield reliable inferences.

21. In the process of setting up the site, they recommended that we hire two or three censors for every 50,000 users. That enables us to back out an estimate of the total number of censors hired within firms at between 50,000 and 75,000, not counting censors within government, 50 Cent Party members, or the Internet police.

22. See (*40*, *41*) for numbers of registered users, which are substantial even if we account for automated sites created by marketing firms (*42*).

23. K. Imai, G. King, E. Stuart, Misunderstandings between experimentalists and observationalists about causal inference. *J. R. Stat. Soc. Ser. A* **171**, 481–502 (2008). doi:10.1111/j.1467-985X.2007.00527.x

24. All posts were made to mainland China accounts. Some were submitted from outside China, when feasible, and many from within China. Recent work has noted that overseas accounts are subject to less stringent censorship regulations than mainland accounts (*43*). This issue does not affect our work because all accounts created and used were mainland China accounts. Users could control account location when creating the account by specifying a location in China, by entering a local mobile number, or by creating the account from a local IP address. We used all of these methods.

25. For each of our three rounds, we wrote 200 posts on non–collective action events (split equally between pro- and anti-government) and 200 posts on collective action events or edge cases (again split equally between pro- and anti-government). Thus, 600 posts submitted relate to non–collective action events, and 600 relate to collective action events or edge cases. We have in total four collective events and two edge cases, and so 400 posts focused on collective action events and 200 on edge cases.

26. J. Crandall, M. Crete-Nishihata, J. Knockel, S. McKune, A. Senft, D. Tseng, G. Wiseman, Chat program censorship and surveillance in China: Tracking TOM-Skype and Sina UC. *First Monday* **18** (no. 7) (July 2013). doi:10.5210/fm.v18i7.4628

27. J. Fallows, The connection has been reset. *Atlantic* (March 2008); www.theatlantic.com/magazine/archive/2008/03/-the-connection-has-been-reset/306650.

28. K. Imai, L. Keele, D. Tingley, T. Yamamoto, Unpacking the black box of causality: Learning about causal mechanisms from experimental and observational studies. *Am. Pol. Sci. Rev.* **105**, 765–789 (2011). doi:10.1017/S0003055411000414

29. J. Pearl, Direct and indirect effects. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence* (Morgan Kaufmann, San Francisco, 2001), pp. 411–420; https://dslpitt.org/uai/papers/01/p411-pearl.pdf.

30. O. Blanchard, A. Shleifer, "Federalism with and without political centralization: China versus Russia" (National Bureau of Economic Research, Cambridge, MA, 2000); www.nber.org/papers/w7616.

31. S. Heilmann, E. Perry, *Mao's Invisible Hand: The Political Foundations of Adaptive Governance in China* (Harvard University Asia Center, Cambridge, MA, 2011).

32. Y. Qian, G. Roland, Federalism and the soft budget constraint. *Am. Econ. Rev.* **88**, 1143–1162 (1998).

33. Y. Qian, B. R. Weingast, Federalism as a commitment to perserving market incentives. *J. Econ. Perspect.* **11**, 83–92 (1997). doi:10.1257/jep.11.4.83

34. G. Egorov, K. Sonin, Dictators and their viziers: Endogenizing the loyalty-competence trade-off. *J. Eur. Econ. Assoc.* **9**, 903–930 (2011). doi:10.1111/j.1542-4774.2011.01033.x

35. G. Salton, *Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer* (Addison-Wesley, Reading, MA, 1988).

36. D. Kelleher, S. Luz, Automatic hypertext keyphrase detection. In *Proceedings of the 19th International Joint Conference on Artificial Intelligence* (Erlbaum, Hillsdale, NJ, 2005), pp. 1608–1609.

37. For examples see (*44*).

38. For news coverage of the protests, see (*45–48*).

39. D. Hopkins, G. King, A method of automated nonparametric content analysis for social science. *Am. J. Pol. Sci.* **54**, 229–247 (2010). doi:10.1111/j.1540-5907.2009.00428.x

40. K. Hong, China's Twitter-like Sina Weibo service now has over 50 million active users per day. *The Next Web*, 13 August 2013; http://tnw.co/1fdNFPS.

41. S. Millward, Tencent Weibo, the 'other weibo' that nobody cares about, reaches 540 million users. *Tech in Asia*, 22 January 2013. http://bit.ly/1byjSNW.

42. K. W. Fu, M. Chau, Reality check for the Chinese microblog space: A random sampling approach. *PLOS ONE* **8**, e58356 (2013). Medline doi:10.1371/journal.pone.0058356

43. J. Q. Ng, Weibo keyword un-blocking is not a victory against censorship. *Tea Leaf Nation*, 21 June 2013; http://bit.ly/1kfqNBC.

44. P. Barefoot, "Principal, get a room with me, spare the schoolchildren!" *China Smack*, 31 May 2013; http://j.mp/19yuv7E.

45. Al-Jazeera, Democracy push as Hong Kong marks handover. 1 July 2013; http://j.mp/145Jvpp.

46. S. Lee, K. Wong, Hong Kong protests to underscore Leung's record-low appeal. *Bloomberg BusinessWeek*, 28 June 2013; http://j.mp/13r3v7v.

47. J. Ngo, July 1 protest is Hong Kong's taste of democracy. *South China Morning Post*, 30 June 2013; http://j.mp/15PcwBt.

48. C. Yung, Annual Hong Kong protest to focus ire on leader. *Wall Street Journal*, 28 June 2013; http://j.mp/13FJB3w.

49. G. King, J. Pan, M. E. Roberts, Replication Data for: Reverse Engineering Chinese Censorship: Randomized Experimentation and Participant Observation (2014). doi:10.7910/DVN/26212

50. K. Imai, L. Keele, D. Tingley, T. Yamamoto, Unpacking the black box of causality: Learning about causal mechanisms from experimental and observational studies. *Am. Pol. Sci. Rev.* **105**, 765–789 (2011). doi:10.1017/S0003055411000414

51. G. King, E. Gakidou, N. Ravishankar, R. T. Moore, J. Lakin, M. Vargas, M. M. Téllez-Rojo, J. E. Hernandez Ávila, M. Hernández Ávila, H. Hernández Llamas, A "politically robust" experimental design for public policy evaluation, with application to the Mexican universal health insurance program. *J. Policy Anal. Manage.* **26**, 479–506 (2007). Medline doi:10.1002/pam.20279