How Censorship in China Allows Government Criticism but Silences Collective Expression*

Gary King† Jennifer Pan‡ Margaret Roberts§

September 24, 2012

Abstract

We offer the first large scale, multiple source analysis of the outcome of what may be the most extensive effort to selectively censor human expression ever implemented. To do this, we have devised a system to locate, download, and analyze the content of millions of social media posts originating from nearly 1,400 different social media services all over China before the Chinese government is able to find, evaluate, and censor (i.e., remove from the Internet) the large subset they deem objectionable. Using modern computer-assisted text analytic methods that we adapt to and validate in the Chinese language, we compare the substantive content of posts censored to those not censored over time in each of 85 topic areas. Contrary to previous understandings, posts with negative, even vitriolic, criticism of the state, its leaders, and its policies are not more likely to be censored. Instead, we show that the censorship program is aimed at curtailing collective action by silencing comments that represent, reinforce, or spur social mobilization, regardless of content. Censorship is oriented toward attempting to forestall collective activities that are occurring now or may occur in the future — and, as such, seem to clearly expose government intent.

*Our thanks to Peter Bol, John Carey, Justin Grimmer, Iain Johnston, Bill Kirby, Jean Oi, Liz Perry, Bob Putnam, Susan Shirk, Noah Smith, Andy Walder, and Barry Weingast for helpful comments and suggestions; the incredible teams at Crimson Hexagon, and the Institute for Quantitative Social Science at Harvard University, for help with data and many technical issues; and our indefatigable undergraduate research associates, Wanxin Cheng, Jennifer Sun, Hannah Waight, Yifan Wu, and Min Yu.

†Albert J. Weatherhead III University Professor, Institute for Quantitative Social Science, 1737 Cambridge Street, Harvard University, Cambridge MA 02138; http://GKing.harvard.edu, king@harvard.edu, (617) 500-7570.

‡Ph.D. Candidate, Department of Government, 1737 Cambridge Street, Harvard University, Cambridge MA 02138; http://people.fas.harvard.edu/∼jjpan/, (917) 740-5726.

§Ph.D. Candidate, Department of Government, 1737 Cambridge Street, Harvard University, Cambridge MA 02138; http://scholar.harvard.edu/mroberts/home

1 Introduction

The size and sophistication of the Chinese government's program to selectively censor the expressed views of the Chinese people is unprecedented in recorded world history. Unlike in the U.S., where social media is centralized through a few providers, in China it is fractured across hundreds of local sites, with each individual site privately employing up to 1,000 censors. Additionally, approximately 20,000–50,000 Internet police and an estimated 250,000–300,000 "50 cent party members" (*wumao dang*) are employed by the central government. However, all levels of government — central, provincial, and local — participate in this huge effort (Chen and Ang 2011, and our interviews with informants, granted anonymity). China overall is tied with Burma at 187th of 197 countries on a scale of press freedom (Freedom House, 2012), but the Chinese censorship effort is by far the largest.

In this paper, we show that this program, designed to *limit* freedom of speech of Chinese people, paradoxically also *exposes* an extraordinarily rich source of information about the Chinese government's interests, intentions, and goals — a subject of long-standing interest to the scholarly and policy communities. The information we unearth is available in continuous time, rather than the usual sporadic media reports of the leaders' sometimes visible actions. We use this new information to develop a theory of the overall purpose of the censorship program, and thus to reveal some of the most basic goals of the Chinese leadership that until now have been the subject of intense speculation but necessarily little empirical analysis. The information we unearth is also a treasure trove that can be used for many other scholarly (and practical) purposes. Upon publication, we will make available replication information for further analyses by other scholars.

Our central theoretical finding is that, contrary to much research and commentary, the purpose of the censorship program is *not* to supress criticism of the state or the Communist Party. Indeed, despite widespread censorship of social media, we find that when the Chinese people write scathing criticisms of their government and its leaders, the probability that their post will be censored does not increase. Instead, we find that the purpose of the censorship program is to reduce the probability of collective action by clipping social

ties whenever any localized social movements are in evidence or expected. We demonstrate these points and then discuss their far-reaching implications for many research areas within the study of Chinese politics and other nondemocratic regimes.

We begin in Section 2 by defining two theories of Chinese censorship. Section 3 describes our unique data source and how we gathered it. Section 4 lays out our strategy for analysis. Section 5 gives our results, and Section 6 concludes.

2 Government Intentions and the Purpose of Censorship

Previous Indicators of Government Intent Deciphering the opaque intentions and goals of the leaders of the Chinese regime was once the central focus of scholarly research on Chinese Communist Party politics, where Western researchers used Kremlinology — or Pekingology — as a methodological strategy (Chang, 1983; Charles, 1966; Hinton, 1955; MacFarquhar, 1974, 1983; Schurmann, 1966; Teiwes, 1979). With the Cultural Revolution and with China's economic opening, more sources of data became available to researchers, and scholars shifted their focus to areas where information was more accessible. Studies of China today rely on government statistics, public opinion surveys, interviews with local officials, as well as measures of the visible actions of government officials and the government as a whole (Guo, 2009; Kung and Chen, 2011; Tsai, 2007*a*,*b*; Shih, 2008). These sources are well-suited to answer other important political science questions, but in gauging government intent, they are widely known to be indirect, very sparsely sampled, and often of dubious value. For example, government statistics, such as the number of protest incidents with government intervention, could offer a view of government interests, but only if we could somehow separate true numbers from government manipulation. Similarly, sample surveys can be informative, but the government obviously keeps information from people, and even when they have the information researchers are seeking respondents may not be willing to express themselves freely. In situations where direct interviews with officials are possible, researchers are in the position of having to read tea leaves to ascertain what their informants really believe.

Measuring intent is all the more difficult with the sparse information coming from

existing methods because the Chinese government is not a monolithic entity. In fact, in those instances when different agencies, leaders, or levels of government work at cross-purposes, even the concept of a unitary intent or motivation may be difficult to define, much less measure. We cannot solve all these problems, but by providing more information about the state's revealed preferences through their censorship behavior, we may be somewhat better able to produce useful measures of intent.

Theories of Censorship We attempt to complement the important work on how censorship is conducted, and how the Internet may increase the space for public discourse (Qiang, 2011; Esarey and Qiang, 2008, 2011; Lindtner and Szablewicz, 2011; Herold, 2011; Yang, 2009; MacKinnon, 2012), by beginning to build an empirically documented theory of why the government censors and what it is trying to achieve through this extensive program. While current scholarship draws the reasonable but broad conclusion that Chinese government censorship is aimed at maintaining the status quo for the current regime, we focus on what specifically the government believes is critical, and what actions it takes, to accomplish this goal.

To do this, we distinguish two theories of what constitutes the goals of the Chinese regime as implemented in their censorship program, each reflecting a different perspective on what threatens the stability of the regime. First is a *state critique* theory, which posits that the goal of the Chinese leadership is to supress dissent, and to prune human expression that finds fault with elements of the Chinese state, its policies, or its leaders. The result is to make the sum total of available public expression more favorable to those in power. Many types of state critique are included in this idea, such as poor government performance.

Second is what we call the theory of *collective action potential*: the target of censorship is people who join together to express themselves collectively, stimulated by someone other than the government, and seem to have the potential to generate collective action. In this view, collective expressions — many people communicating on social media on the same subject — regarding actual collective actions, such as protests, as well as those about events that seem likely to generate collective actions but have not yet done so, are likely

to be censored. Whether social media posts with collective action potential find fault with or assign praise to the state, or are about subjects unrelated to the state, is unrelated to this theory.

An alternative way to describe what we call "collective action potential" is the apparent perspective of the Chinese government, where collective expression organized outside of governmental control equals factionalism and ultimately chaos and disorder. For example, on the eve of Communist Party's 90th birthday, the state-run Xinhua news agency issued an opinion that western-style parliamentary democracy would lead to a repetition of the turbulent factionalism of China's Cultural Revolution (http://j.mp/McRDXk). Similarly, at the Fourth Session of the 11th National Peoples Congress in March of 2011, Wu Bangguo, member of the Politburo Standing Committee and Chairman of the Standing Committee of the National People's Congress, said that "On the basis of China's conditions... we'll not employ a system of multiple parties holding office in rotation" in order to avoid "an abyss of internal disorder" (http://j.mp/Ldhp25). China observers have often noted the emphasis placed by the Chinese government on maintaining stability (Shirk, 2007; Whyte, 2010; Zhang et al., 2002), as well as the government's desire to limit collective action by clipping social ties (Perry, 2002, 2008). The Chinese regime encounters a great deal of contention and collective action; according to Sun Liping, a professor of Sociology at Tsinghua University, China experienced 180,000 "mass incidents" in 2010 (http://j.mp/McQeji). Because the government encounters collective action frequently, it influences the actions and perceptions of the regime. The stated perspective of the Chinese government is that limitations on horizontal communications is a legitimate and effective action designed to protect its people (Perry, 2010) — in other words, a paternalistic strategy to avoid chaos and disorder, given the conditions of Chinese society.

Current scholarship has not been able to differentiate empirically between the two theories we offer. Marolt (2011) writes that online postings are censored when they "either criticize China's party-state and its policies directly or advocate collective political action." MacKinnon (2012) argues that during the Wenzhou high speed rail crash, Internet content providers were asked to "track and censor critical postings." Esarey and Qiang

(2008) find that Chinese bloggers use satire to convey criticism of the state in order to avoid harsh repression. Esarey and Qiang (2011) write that party leaders are most fearful of "Concerted efforts by influential netizens to pressure the government to change policy," but identify these pressures as criticism of the state. Shirk (2011) argues that the aim of censorship is to constrain the mobilization of political opposition, but her examples suggest that critical viewpoints are those that are suppressed.

Collective action in the form of protests is often thought to be the death knell of authoritarian regimes. Protests in East Germany, Eastern Europe, and most recently the Middle East have all preceded regime change (Ash, 2002; Lohmann, 1994; Przeworski et al., 2000). A great deal of scholarship on China has focused on what leads people to protest and their tactics (Blecher, 2002; Cai, 2002; Chen, 2000; Lee, 2007; O'Brien and Li, 2006; Perry, 2002, 2008). The Chinese state seems focused on preventing protest at all costs—and, indeed, the prevalence of collective action is part of the formal evaluation criteria for local officials (Edin, 2003). However, several recent works argue that authoritarian regimes may expect and welcome substantively narrow protests as a way of enhancing regime stability by identifying, and then dealing with, discontented communities (Dimitrov, 2008; Lorentzen, 2010; Chen, 2012). Chen (2012) argues that small, isolated protests have a long tradition in China and are an expected part of government.

Outline of Results The nature of the two theories means that either or both could be correct or incorrect. Here, we offer evidence that, with few exceptions, the answer is simple: state critique theory is incorrect and the theory of collection action potential is correct. Our data show that the Chinese censorship program allows for a wide variety of criticisms of the Chinese government, its officials, and its policies. As it turns out, censorship is primarily aimed at restricting the spread of information that may lead to collective action, regardless of whether or not the expression is in direct opposition to the state and whether or not it is related to government policies. Large increases in online volume are good predictors of censorship when these increases are associated with events related to collective action, e.g., protest on the ground. In addition, we measure sentiment within each of these events and show that during these events, the government censors views

that are both supportive and critical of the state. These results reveal that the Chinese regime believes supressing social media posts with collective action potential, rather than suppression of criticism, is crucial to its maintaining power.

3 Data

We describe here the challenges involved in collecting large quantities of detailed information that the Chinese government does not want anyone to see and goes to great lengths to prevent anyone from accessing. We discuss the types of censorship we study, our data collection process, the limitations of this study, and ways we organize the data for subsequent analyses.

3.1 Types of Censorship

Human expression is censored in Chinese social media in at least four ways, the last of which is the focus of our study. First is "The Great Firewall of China," which disallows certain entire web sites from operating in the country. The Great Firewall is an obvious problem for foreign Internet firms, and for the Chinese people interacting with others outside of China on these services, but it does little to limit the expressive power of Chinese people who can find other sites to express themselves in similar ways. For example, Facebook is blocked in China but RenRen is a close substitute; similarly Sina Weibo is a popular Chinese clone of Twitter, which is also unavailable.

Second is "keyword blocking" which stops a user from posting text that contain banned words or phrases. This has only a minor effect on freedom of speech, since netizens can easily outwit automated programs. To do so, they use analogies, metaphors, satire, and other evasions. The Chinese language offers novel evasions, such as substituting characters for those banned with others that have unrelated meanings but sound alike ("homophones") or look similar ("homographs"). An example of a homograph is 目田, which has the nonsensical literal meaning of "eye-field" but is used by World of Warcraft players to substitute for the banned but similarly shaped 自由 which means freedom. As an example of a homophone, the sound "hexie" is often written as 河蟹, which means

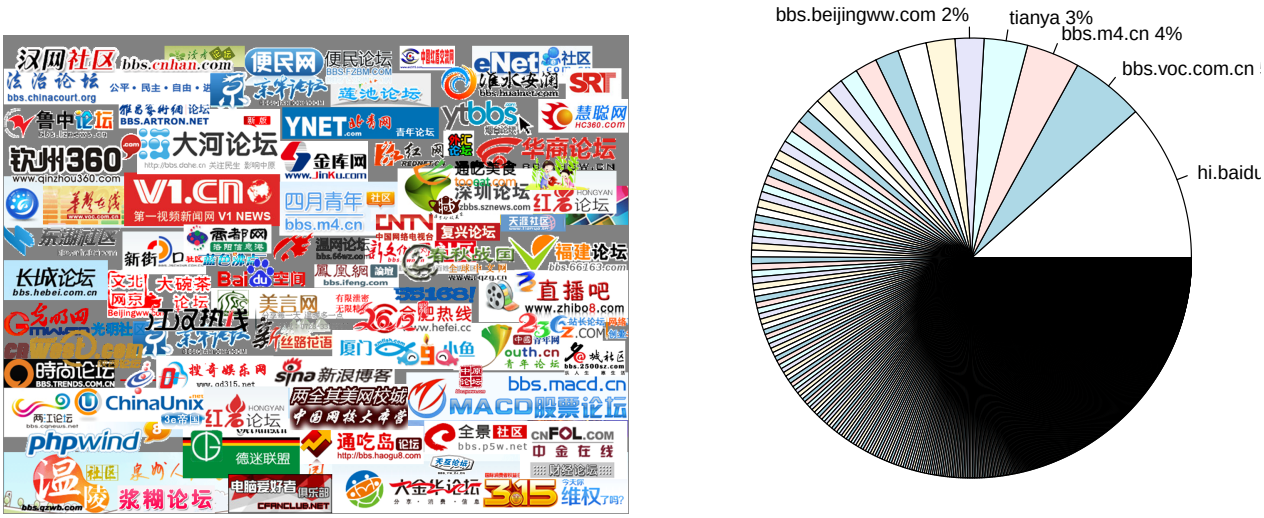
"river crab," but is used to refer to 和谐, which is the official state policy of a "harmonious society".

Once past the first two barriers to freedom of speech, the text gets posted on the web and the censors read and remove those they find objectionable. Near as we can tell from the literature, observers, private conversations with those inside several governments, and an examination of the data, the censors accomplish their task almost entirely by reading each post by hand. Automated methods appear to be a minor or nonexistent part of this effort, perhaps because the state of methods of automated classification is not up to the task and because labor is inexpensive. Unlike The Great Firewall and keyword blocking, hand censoring cannot be evaded by clever phrasing: anything obscure enough to evade the censors would also likely evade most of the audience as well. Thus, it is this last and most extensive form of censoring that we focus on in this paper.

3.2 Collection

We begin with social media blogs in which it is at least possible for writers to express themselves fully, prior to possible censorship, and leaving to other research social media services that constrain authors to very short Twitter-like (*weibo*) posts (e.g., Bamman, O'Connor and Smith, 2012). In many countries, such as the U.S., almost all blog posts appear on a few large sites (Facebook, Google's blogspot, Tumblr, etc.); China does have some big sites such as sina.com, but a large portion of its social media landscape is finely distributed over numerous individual sites, e.g., local bbs forums. This difference poses a considerable logistical challenge for data collection — with different web addresses, different software interfaces, different companies and local authorities monitoring those accessing the sites, different network reliabilities, access speeds, terms of use, and censorship modalities, and different ways of potentially hindering or stopping our data collection. Fortunately, the structure of Chinese social media also turns out to pose a special opportunity for studying localized control of collective expression, since the numerous local sites provide considerable information about the geolocation of posts, much more than is available even then in the U.S.

The most complicated engineering challenges in our data collection process involves



(a) Sample of Sites

(b) All Sites excluding Sina

Figure 1: The Fractured Structure of the Chinese Social Media Landscape

locating, accessing, and downloading posts from many web sites before Internet content providers or the government reads and censors those that are deemed by authorities as objectionable;[1] revisiting each post frequently enough to learn if and when it was censored; and proceeding with data collection in so many places in China without affecting the system we were studying or being prevented from studying it. The reason we are able to accomplish this is because our data collection methods are highly automated whereas Chinese censorship is a massive effort accomplished largely by hand. Our extensive engineering effort, which we do not detail here for obvious reasons, is executed at many locations around the world, including inside China.

Ultimately, we were able to locate, obtain access to, and download social media posts from 1,382 Chinese websites during the first half of 2011. The most striking feature of the structure of Chinese social media is its extremely long (power-law like) tail. Figure 1 gives a sample of the sites and their logos in Chinese (in panel a) and a pie chart of the number of posts that illustrate this long tail (in panel b). The largest sources of posts include blog.sina (with 59% of posts), hi.baidu, voc, bbs.m4, and tianya, but the tail keeps going.[2]

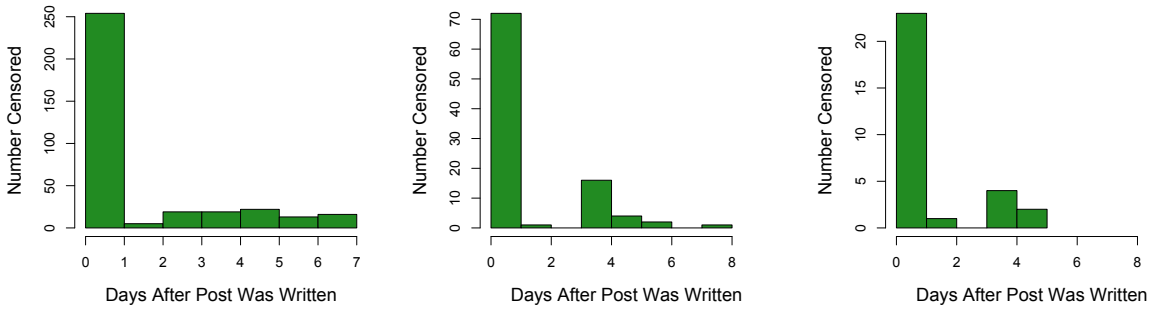
[1] See MacKinnon (2012) for additional information on the censorship process.

[2] See http://blog.sina.com.cn/, http://hi.baidu.com/, http://voc.com.cn/, http://bbs.m4.cn/, and http://tianya.cn/.

Social media posts cover such a huge range of topics that a random sampling strategy attempting to cover everything is rarely informative about any individual topic of interest. Thus, we begin with a stratified random sampling design, organized hierarchically. We first choose 85 separate topic areas within three categories of hypothesized political senstivity, ranging from "High" (such as Ai Weiwei) to "Medium" (such as the one child policy) to "Low" (such as a popular online video game). We chose the specific topics within these categories by reviewing prior literature, consulting with China specialists, and studying current events. Appendix A gives a complete list. Then, within each topic area, defined by a set of keywords, we collected all social media posts over a six month period. We examined the posts in each area, removed spam, and explored the content with the tool for computer-assisted reading (Grimmer and King, 2011; Crosas et al., 2012). With this procedure we collected 3,674,698 posts, with 127,283 randomly selected for further analysis. (We repeated this procedure for other time periods, and in some cases in more depth for some issue areas, and overall collected and analyzed 11,382,221 posts.) All posts originated from sites in China, were written in Chinese, and excluded those from Hong Kong and Taiwan. For each post, we examined its content, placed it on a timeline according to topic area, and revisited the website from which it came repeatedly thereafter to determine whether it was censored. We supplemented this information with other specific data collections as needed.

The censors are not shy, and so we found it straightforward to distinguish (intentional) censorship from sporadic outages or transient time-out errors. The censored web sites include notes such as "Sorry, the host you were looking for does not exist, has been deleted, or is being investigated" (抱歉，指定的主题 不存在或已被删除或正在被审核) and are sometimes even adorned with pictures of Jingjing, an Internet police cartoon character.

Although our methods are faster than the Chinese censors, the censors nevertheless appear highly expert at their task. We illustrate this with analyses of random samples of posts surrounding the 9/27/2011 Shanghai Subway crash, and posts collected between 4/10/2012 and 4/12/2012 about Bo Xilai, a recently deposed member of the Chinese elite, and a separate collection of posts about his wife, Gu Kailai, who was accused of murder.



(a) Shanghai Subway Crash (b) Bo Xilai (c) Gu Kailai

Figure 2: The Speed of Censorship, Monitored in Real-Time

We monitored each of the posts in these three areas continuously in near real time for 9 days. (Censorship in other areas follow the same basic pattern.) Histograms of the time until censorship appear in Figure 2. For all three, the vast majority of censorship activity occurs within 24 hours of the original posting, although a few deletions occur longer than five days later. This is a remarkable organizational accomplishment, requiring large scale military-like precision: The many leaders at different levels of government and at different internet content providers first need to come to a decision (by agreement, direct order, or compromise) about what to censor in each situation; they need to communicate it to tens or hundreds of thousands of individuals; and then they must all complete execution of the plan within about 24 hours. As Edmond (2012) points out, the proliferation of information sources on social media makes information more difficult to control; however, the Chinese government has overcome these obstacles on a national scale. Given the normal human difficulties of coming to agreement with many others, and the usual difficulty of achieving high levels of inter-coder reliability on interpreting text (e.g., Hopkins and King, 2010, Appendix), the effort the government puts into its censorship program is large, and highly professional. We have found some evidence of disagreements within this large and multifarious bureaucracy, such as at different levels of government, but we have not yet studied these differences in detail.

3.3 Limitations

As we show below, our methodology reveals a great deal about the goals of the Chinese leadership, but it misses self-censorship and censorship that may occur before we are able to obtain the post in the first place; it also does not quantify the direct effects of The Great Firewall, keyword blocking, or search filtering in finding what others say. We have also not studied the effect of physical violence, such as the arrest of bloggers, or threats of the same. Although many officials and levels of government have a hand in the decisions about what and when to censor, our data only sometimes enable us to distinguish among these sources.

We are of course unable to determine the consequences of these limitations, although it is reasonable to expect that the most important of these are physical violence and threats and the resulting self-censorship. Although the social media data we analyze include expressions by millions of Chinese and cover an extremely wide range of topics and speech behavior, the presumably much smaller number of discussions we cannot observe are likely to be those of the most (or most urgent) interest to the Chinese government.

Finally, in the past, studies of Internet behavior were judged based on how well their measures approximated "real world" behavior; subsequently, online behavior has become such a large and important part of human life that the expressions observed in social media is now important in its own right, regardless of whether it is a good measure of non-Internet freedoms and behaviors. But either way, we offer little evidence here of connections between what we learn in social media and press freedom or other types of human expression in China.

4 Analysis Strategy

Overall, an average of approximately 13% of all social media posts are censored. This average level is quite stable over time when aggregating over all posts in all areas, but masks enormous changes in volume of posts and censorship efforts. Our first hint of what might (not) be driving censorship rates is a surprisingly low correlation between our ex ante measure of political sensitivity and censorship: Censorship behavior in the Low

and Medium categories was essentially the same (16% and 17% respectively) and only marginally lower than the High category (24%).[3] Clearly something else is going on. To convey what this is, we now discuss our coding rules, our central hypothesis, and the exact operational procedures the Chinese government may use to censor.

4.1 Coding Rules

We discuss our coding rules in five steps. First, we begin with social media posts organized into the 85 topic areas defined by keywords during from our stratified random sampling plan. Although we have conducted extensive checks that these are accurate (by reading large numbers and also via modern computer-assisted reading technology), our topic areas will inevitably (with any machine or human classification technology) include some posts that do not belong. We take the conservative approach of first drawing conclusions even when affected by this error. Afterward, we then do numerous checks (via the same techniques) after the fact to ensure we are not missing anything important. We report below the few patterns that could be construed as a systematic error; each one turns out to strengthen our conclusions.

Second, conversation in social media within almost all topic areas (and countries) is well known to be highly "bursty," that is with periods of stability punctuated by occasional sharp spikes in volume around specific subjects (Ratkiewicz et al., 2010). We also found that with only two exceptions — pornography and criticisms of the censors, described below — censorship effort is often especially intense within *volume bursts*. Thus, we organize our data around these volume bursts. We think of each of the 85 topic areas as a six month time series of daily volume and detect bursts using the weights calculated from robust regression techniques to identify outlying observations from the rest of the time series (Huber, 1964; Rousseeuw and Leroy, 1987). In our data, this sophisticated burst detection algorithm is almost identical to using time periods with volume more than three standard deviations greater than the rest of the six month period. With this procedure, we detected 86 distinct volume bursts within 66 of the 85 topic areas.[4]

[3] That all three figures are higher than the average level of 13% reflects the fact that the topic areas we picked ex ante had generated at least some public discussion.

[4] We attempted to identify duplicate posts, the Chinese equivalent of "retweets", sblogs (spam blogs),

Third, we examined the posts in each volume burst and identified the real world *event* associated with the online conversation. This was easy and the results unambiguous.

Fourth, we classified each event into one of five content areas: (1) collective action potential, (2) criticism of the censors, (3) pornography, (4) government policies, and (5) other news. As with topic areas, each of these categories may include posts that are critical or not critical of the state, its leaders, and its policies. We define collective action as the pursuit of goals by more than one person controlled or spurred by actors other than government officials or their agents. Our theoretical category of "collective action potential" involves any event that has the potential to cause collective action, but to be conservative, and to ensure clear and replicable coding rules, we limit this category to events on topics which (a) involve protest or organized crowd formation outside the Internet; (b) individuals who have organized or incited collective action on the ground in the past; or (c) topics related to nationalism or nationalist sentiment that have incited protest or collective action in the past. (Nationalism is treated separately because of its frequently demonstrated high potential to generate collective action and also to constrain foreign policy, an areas which has long been viewed as a special prerogative of the government; Reilly 2012.)

Events are categorized as criticism of censors if they pertain to government or non-government entities with control over censorship, including individuals and firms. Pornography includes advertisements and news about movies, websites, and other media containing pornographic or explicitly sexual content. Policies refer to government statements or reports of government activities pertaining to domestic or foreign policy. And "other news" refers to reporting on events, other than those which fall into one of the other four categories.

Finally, we conducted a study to verify the reliability of our event coding rules. To do this, we gave our rules above to two people familiar with Chinese politics and asked them to code each of the 86 events (each associated with a volume burst) into one of the five categories. The coders worked independently and classified each of the events on their own. Decisions by the two coders agreed in 98.8% (i.e., 85 of 86) of the events. The only event with divergent codes was the pelting of Fang Binxing (the architect of China's Great

and the like. Excluding these posts had no noticeable effect on our results.

Firewall) with shoes and eggs. This event included criticism of the censors and to some extent collective action because several people were working together to throw things at Fang. We broke the tie by counting this event as an example of criticism of the censors, but however this event is coded does not affect our results since we predict both will be censored.

4.2 Central Hypothesis

Our central hypothesis is that the government censors *all* posts in topic areas during volume bursts that discuss events with collective action potential. That is, *the censors do not judge whether individual posts have collective action potential*, perhaps in part because rates of intercoder reliability would likely be very low. In fact, Kuran (1989) and Lohmann (2002) show that it is information about a collective action event that propels collective action and so distinguishing this from explicit calls for collective action is difficult if not impossible. Instead, we hypothesize that they make the much easier judgment, about whether the posts are on topics associated with events that have collective action potential, and they do it regardless of whether they criticize the state or not.

The censors also attempt to censors all posts in the categories of pornography and criticism of the censors, but not within event categories of government policies and news.

4.3 The Government's Operational Procedures

The exact operational procedures by which the Chinese government censors is of course not observed. Our coding rules can be viewed as our attempt at an approximation to them. We define topic areas by hand, sort social media posts into topic areas by keywords, and detect volume bursts automatically via statistical methods for time series data on post volume. These steps might be combined by the government to detect topics automatically based on spikes in posts with high similarity, but this would likely involve considerable error given inadequacies in fully automated clustering technologies. In some cases, identifying the real world event might occur before the burst, such as if the censors are secretly warned about an upcoming event (such as the imminent arrest of a dissident) that could spark collective action. Identifying events from bursts that were observed first would need

to be implemented at least mostly by hand, perhaps with some help from algorithms that identify statistically improbable phrases. Finally, the actual decision to censor an individual post — which according to our hypothesis involves checking whether it is associated with a particular event — is almost surely accomplished largely by hand, since no known statistical or machine learning technology can achieve a level of accuracy anywhere near that which we observe in the Chinese censorship program. Here, censors may begin with many keyword searches on the event identified, but will need to manually read through the resulting posts to censor those which are related to the triggering event. For example, when censors identified protests in Zengcheng as precipitating online discussion, they may have conducted a keyword search among posts for Zengcheng, but they will have read through these posts by hand to separate posts about protests from posts talking about Zengcheng in other context, say Zengchengs lychee harvest.

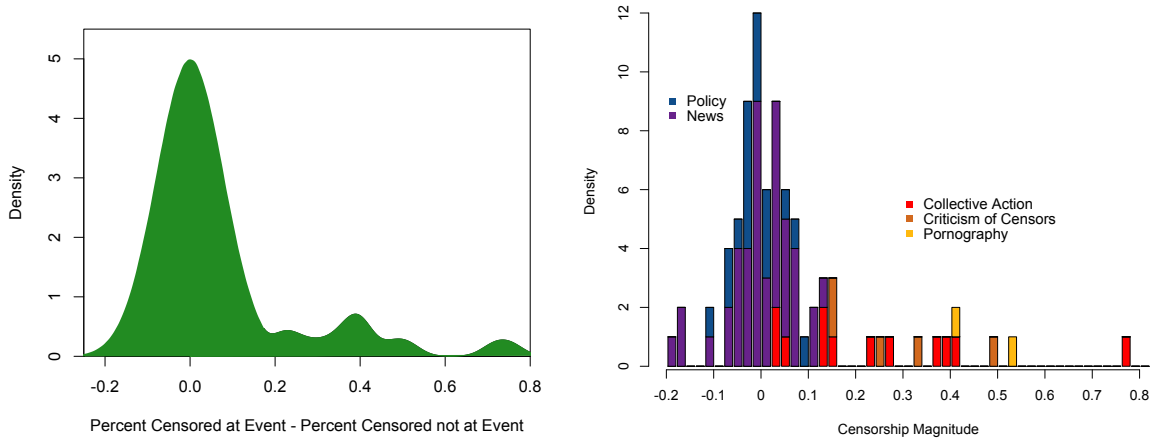
5 Results

We now offer three increasingly specific tests of our hypotheses. These tests are based on (1) post volume, (2) the nature of the event generating each volume burst, and (3) the specific content of the censored posts.

5.1 Post Volume

If the goal of censorship is to stop discussions with collective action potential, then we would expect more censorship during volume bursts than at other times. We also expect some bursts — those with collective action potential — to have much higher levels of censorship.

To begin to study this pattern, we define *censorship magnitude* for a topic area as the percent censored within a volume burst minus the percent censored outside all bursts. (The base rates, which vary very little across issue areas and which we present in detail in graphs below, do not impose empirically relevant ceiling or floor effects on this measure.) This is a stringent measure of the interests of the Chinese government because censoring during a volume burst is obviously more difficult owing to there being more posts to



(a) Distribution of Censorship Magnitude

(b) Censorship Magnitude by Event Type

Figure 3: "Censorship Magnitude," the percent of posts censored inside a volume burst minus outside volume bursts.

evaluate, less time to do it in, and little or no warning of when the event will take place.

Panel (a) in Figure 3 gives a histogram with results that appear to support the hypothesis. The results show that the bulk of volume bursts have a censorship magnitude centered around zero, but with an exceptionally long right tail (and no corresponding long left tail). Clearly volume bursts are often associated with dramatically higher levels of censorship even compared to the baseline during the rest of the six months for which we observe a topic area.

5.2 The Nature of Events Generating Volume Bursts

We now show that volume bursts generated by events pertaining to collective action, criticism of censors, and pornography are censored, albeit as we show in different ways, while post volume generated by discussion of government policy and other news are not. We discuss the state critique hypothesis in the next subsection. Here, we offer three separate, and increasingly detailed, views of our present results.

First, consider Panel (b) of Figure 3, which takes the same distribution of censorship magnitude as in Panel (a) and displays it by event type. The result is dramatic: Collective action, criticism of the censors, and pornography (in red, orange, and yellow) fall largely

to the right, indicating high levels of censorship magnitude, while policies and news fall to the left (in blue and purple). On average, censorship magnitude is 27% for collective action, but -1% and -4% for policy and news. [5]

Second, we list the specific events with the higest and lowest levels of censorship magnitude. These appear, using the same color scheme, in Figure 4. The events with the highest collective action potential include protests in Inner Mongolia and Zengcheng, the arrest of artist-slash-political dissident Ai Weiwei, and the bombings over land claims in Fuzhou. Notably, one of the highest "collective action potential" events was not political at all: following the Japanese earthquake and subsequent meltdown of the nuclear plant in Fukushima, a rumour spread through Zhejiang province that the iodine in salt would protect people from radiation exposure, and a mad rush to buy salt ensued. The rumor was biologically false, and had nothing to do with the state one way or the other, but it was highly censored; the reason appears to be because of the localized control of collective expression by actors other than the government. Indeed, we find that salt rumors on local websites are much more likely to be censored than salt rumors on national websites.[6]

Consistent with our theory of collective action potential, some of the most highly censored events are not criticisms or even discussions of national policies, but rather highly localized collective expressions that represent or threaten group formation. One such example is posts on a local Wenzhou website expressing support for Chen Fei, a environmental activist who supported an environmental lottery to help local environmental protection. Even though Chen Fei is *supported* by the central government, all posts supporting him on the local website are censored, likely because of his record of organizing collective action. In the mid-2000s, Chen founded an environmental NGO (色环保志愿者协会) with more than 400 registered members who created China's first "no-plastic bag village", which eventually led to legislation on use of plastic bags. Another example is a heavily censored group of posts expressing collective anger about lead poisoning in Jiangsu Province's Suyang County from battery factories. These posts talk about children

[5] The baseline (the percent censorship outside of volume bursts) is typically very small, 3-5% and varies relatively little across topic areas.

[6] As in the two relevant events in Figure 4, pornography often appears in social media in association with the discussion of some other popular news or discussion, to attract viewers.

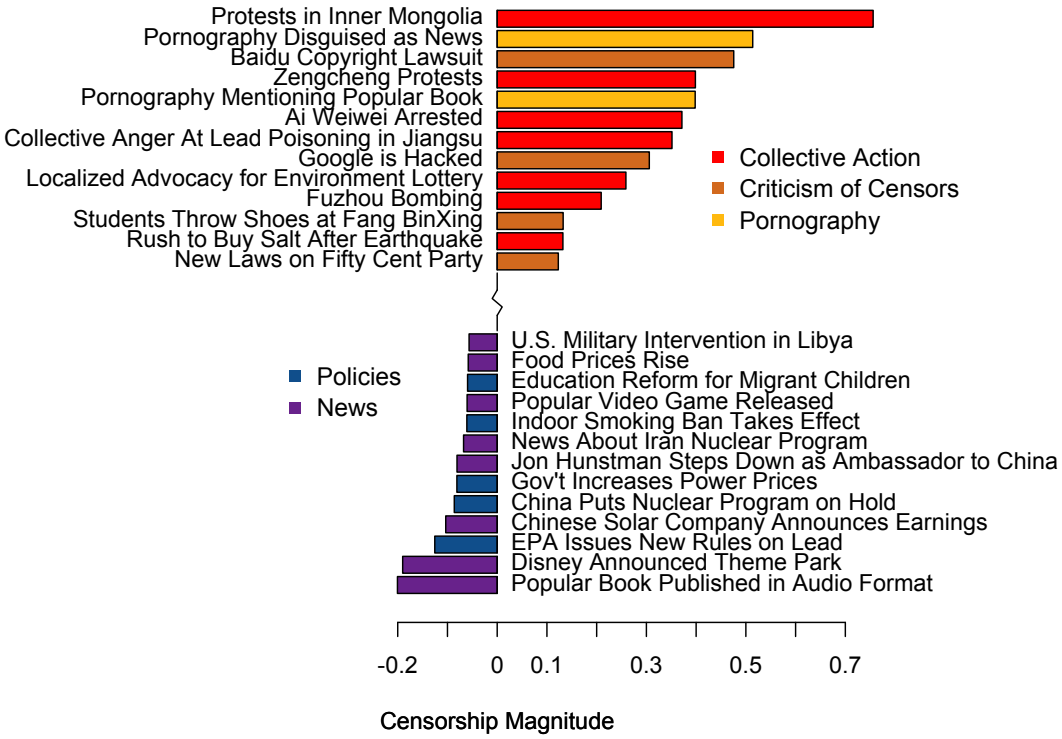


Figure 4: Events with Highest and Lowest Censorship Magnitude

sickened by pollution from lead acid battery factories in Zhejiang province belonging to the Tianneng Group (天能集团), and report that hospitals refused to release results of lead tests to patients. In January 2011, villagers from Suyang gathered at the factory to demand answers. Such collective organization is not tolerated by the censors, regardless of whether it supports the government or criticizes it.

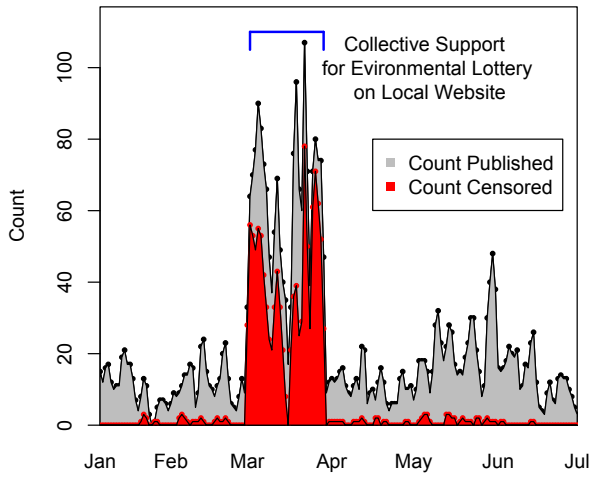
In *all* events categorized as having collective action potential, censorship within the event is more frequent than censorship outside the event. In addition, these events are, on average, considerably more censored than other types of events. These facts are consistent with our theory that the censors are intentionally searching for and taking down posts based on collective action potential. However, we add to these tests one based on an examination of what might lead to different levels of censorship among events within this category: Although we have developed a quantitative measure, some of the events in this category clearly have more collective action potential than others. By studying the specific events, it is easy to see that events with the lowest levels of censorship magnitude

generally have less collective action potential than the very highly censored cases, as consistent with our theory.

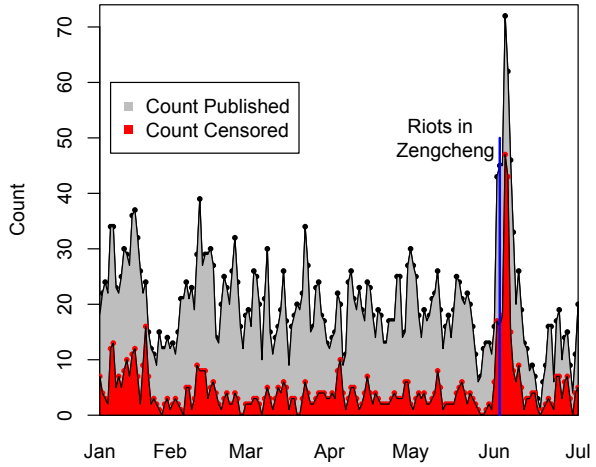
To see this, consider the few events classified as collective action potential with the lowest levels of censorship magnitude. These include a volume burst associated with protests about ethnic stereotypes in the animated children's movie *Kungfu Panda*, which was properly classified as a collective action event, but its potential for future protests is obviously highly limited. Another example is Qian Yunhui, a village leader in Zhejiang, who led villagers to petition local governments for compensation for land seized and was then (supposedly accidentally) crushed to death by a truck. These two events involving Qian had high collective action potential, but both were before our observation period. In our period, there was an event that led to a volume burst around the much narrower and far less incendiary issue of how much money his family was given as a reparation payment for his death.

Finally, we give some more detailed information of a few examples of three types of events, each based on a random sample of posts in one topic area. First, Figure 5 gives four time series plots that initially involve low levels of censorship, followed by a volume spike during which we witness very high levels of censorship. Censorship in these examples are high in terms of the absolute number of censored posts and the percent of posts that are censored. The pattern in all four graphs (and others we do not show) is evident: the Chinese authorities disproportionately focus considerable censorship efforts during volume bursts.

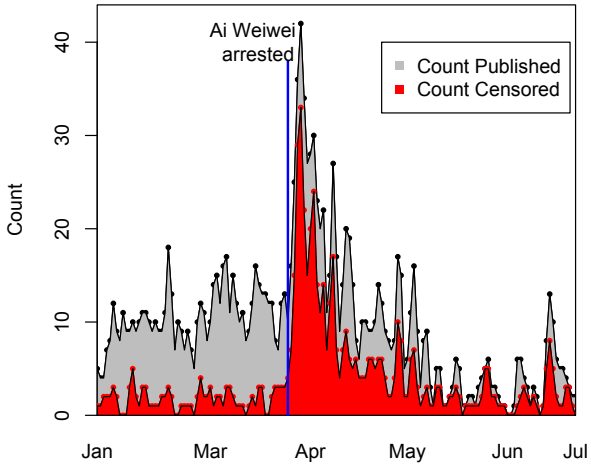
We also went further and analyzed (by hand and via computer-assisted methods described in Grimmer and King 2011) the smaller number of uncensored posts during volume bursts associated with events that have collection action potential, such as in Panel (a) of Figure 5 where the red area does not entirely cover the gray during the volume burst. In this event, and the vast majority of cases like this one, uncensored posts are not about the event, but just happen to have the keywords we used to identify the topic area. Again we find that the censors are highly accurate and aimed at increasing censorship magnitude. Automated methods of individual classification are not capable of this high a



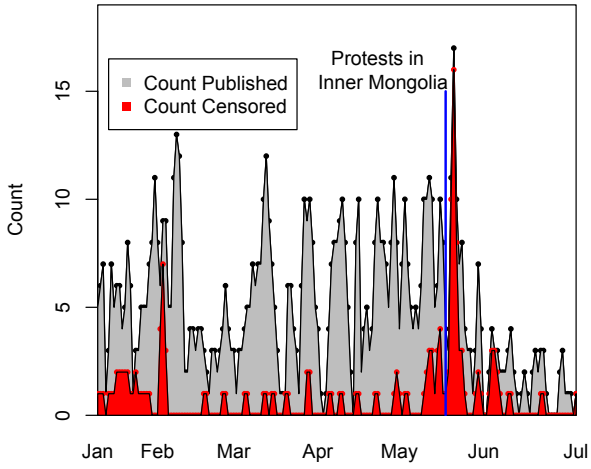
(a) Rush to buy salt after earthquake



(b) Riots in Zengcheng



(c) Dissident Ai Weiwei



(d) Inner Mongolia Protests

Figure 5: High Censorship During Collective Action Events (in 2011)

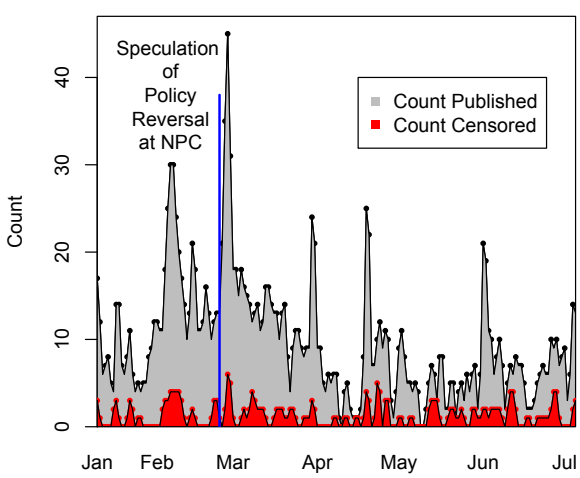
level of accuracy.

Second, we offer four time series plots of random samples of posts in Figure 6 which illustrate topic areas with one or more volume bursts but without censorship. These cover important, controversial, and potentially incendiary topics — including policies involving the one child policy, education policy, and state corruption, as well as news about power prices — but none of the volume bursts where associated with any localized collective expression, and so censorship remains consistently low.

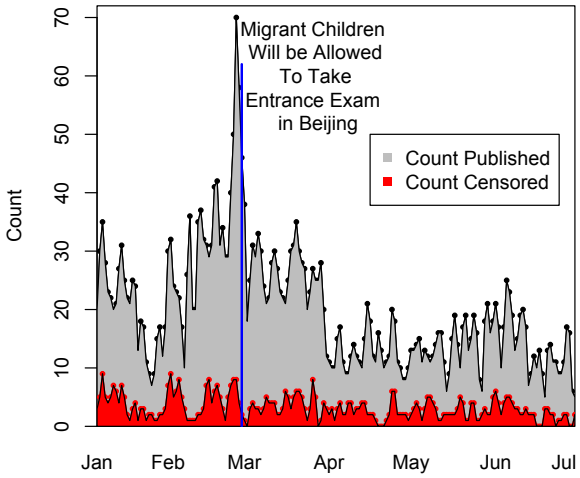
Finally, we found that 83 of 85 topic areas fall into the patterns portrayed by Figures 5 and 6. The two with divergent patterns can be seen in Figure 7. These topics involve analyses of random samples of posts in the areas of pornography (panel a) and criticism of the censors (panel b). What is distinctive about these topics compared to the remaining 83 we studied is that censorship levels remain high consistently the entire six month period and, consequently, do not increase further during volume bursts. Similar to American politicians who talk about pornography as undercutting the "moral fiber" of the country, Chinese leaders describe it as violating public morality and damaging the health of young people, as well as promoting disorder and chaos; regardless, censorship in one form or another is often the consequence.

More striking is an oddly "inappropriate" behavior of the censors: They offer freedom to the Chinese people to criticize every political leader except for the censors, every policy except the one they implement, and every program except the one they run. Even within the strained logic the Chinese state uses to justify censorship, Figure 7 (Panel b) — which reveals consistently high levels of censored posts that involve criticisms of the censors — is remarkable.

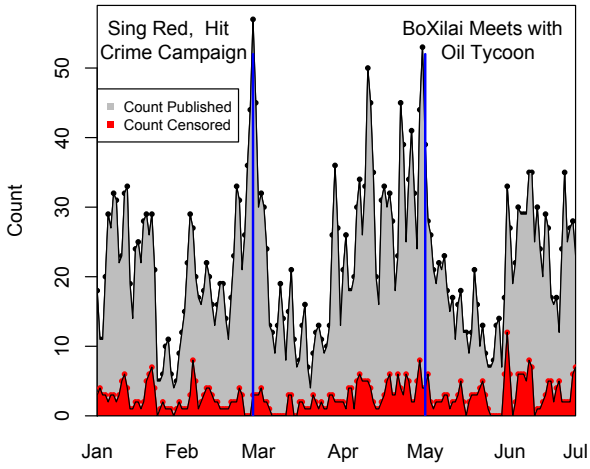
We also went a step further for both analyses in Figure 7 and studied the remaining uncensored posts in both categories. We learned that these are posts, which happened to use the same keywords that we used to define the topic area, turned out to not be substantively relevant. Thus, the posts in Panel (a) which were uncensored were not pornographic (Justice Potter Stewart was right: you really do know it when you see it!); instead, the words we used to select pornographic posts were a homophone that are sometimes used



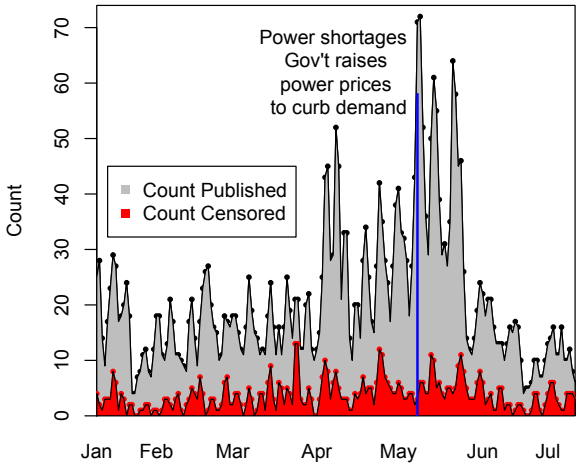
(a) One Child Policy



(b) Education Policy

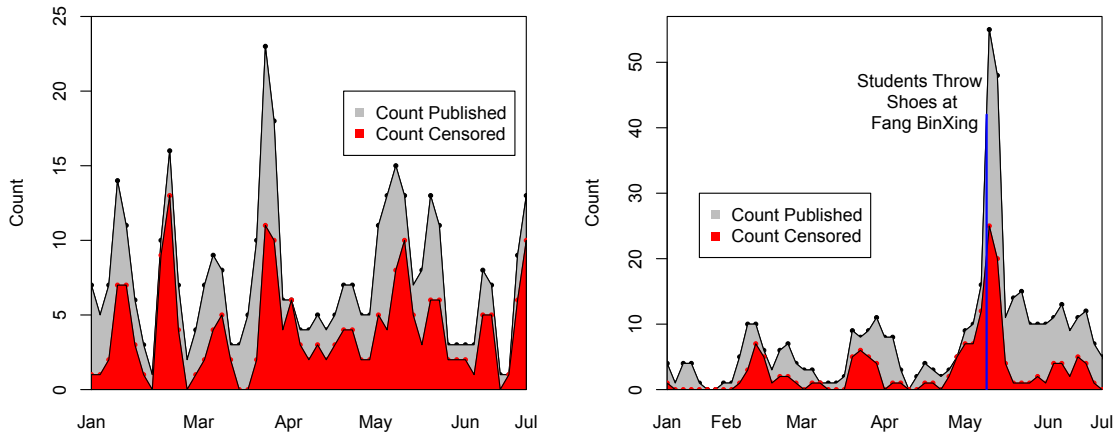


(c) Corruption Policy (Bo Xiali)



(d) News on Power Prices

Figure 6: Low Censorship on News and Policy Events (in 2011)



(a) Pornography

(b) Criticism of the Censors

Figure 7: Two Topics with Continuous High Censorship Levels (in 2011)

for poetic phrasing about unrelated subjects. Similarly, the uncensored posts in Panel (b) mentioned, but did not involve criticism of, the censors. If we redid this figure using this more precise measure, almost no gray would be visible from behind the red wall of censorship; to be conservative, that is in order that we follow same rules set we ex ante, we left them as is.

5.3 Content of Censored and Uncensored Posts

Our final test involves comparing the content of censored and uncensored posts. State critique theory predicts that posts critical of the state are those censored, regardless of their collective action potential. In contrast, the theory of collective action potential predicts that posts related to collective action will be censored regardless of whether they criticize or praise the state, with both critical and supportive posts not censored in the absence of collective action potential.

To conduct this test in a very large number of posts, we need a method of automated text analysis that can accurately estimate the percentage of posts in each category of any given categorization scheme. We thus adapt to the Chinese language the methodology introduced in the English language by Hopkins and King (2010). This method does not

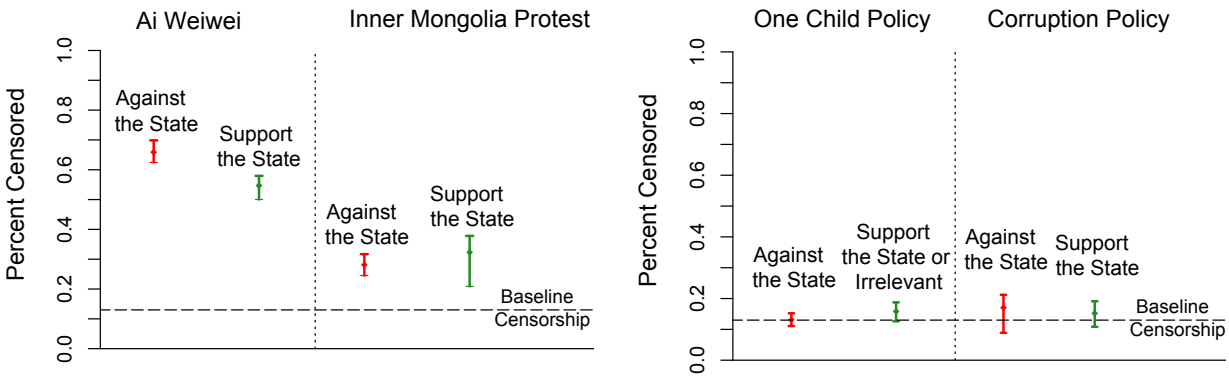
require (inevitably error prone) machine translation, individual classification algorithms, or identification of a list of keywords associated with each category; instead, it requires a small number of posts to be read and categorized in the original Chinese. We conducted a series of rigorous validation tests and obtain highly accurate results — as accurate as if it were possible to read and code all the posts by hand, which of course is not feasible. We describe these procedures, and give a sample of our validation tests, in Appendix B.

For our analysis, we use categories of posts that are (1) against the state, (2) for the state, or (3) irrelevant or factual reports about the events. However, we are not interested in the percent of posts in each of these categories, which would be the usual output of the Hopkins and King procedure. We are also not interested in the percent of posts in each category among those posts which were censored and among those which were not censored, which would result from running the Hopkins-King procedure once on each set of data. Instead, we need to estimate and compare the percent of posts censored in each of the three categories. Appendix B thus also shows how to use use Bayesian logic to extend the Hopkins-King procedure to our quantities of interest.

We begin by analyzing two of the high collective action events covered in Figure 5 — the arrest of Ai Weiwei and protests in Inner Mongolia. As a harder test, we study all posts within the six month period covered by each of these topic areas, rather than the less diverse posts only within each volume burst. Panel (a) of Figure 8 gives the percent of posts censored. As is clear, posts that are against the state (in red) or for the state (in green) are *both* censored at a high and very similar level, considerably above the baseline censorship level. A hypothesis test indicates no significant difference between the two categories in percent censorship for each event.[7] This clearly shows support for the collective action potential theory and against the state critique theory of censorship.

We also conduct a parallel analysis for two topics, taken from the analysis in Figure 6, that cover policies without collective action potential events — one child policy and corruption policy. In this situation, we again get the empirical result that is consistent with our theory, in both analyses: Categories against and for the state both fall at about

[7] A hypothesis test for the difference in two quantities involves the uncertainty reflected in the error bars in the plot as well as the covariance between the two.



(a) Collective Action Events

(b) Non-Collective Action Events

Figure 8: Content of Censored Posts

the same, baseline level of censorship.

The results are clear: posts are censored if they are in a topic area with collective action potential and not otherwise. Whether or not the posts are in favor of the government, its leaders, and its policies has no effect on the probability of censorship.

We conclude this section with some examples of posts to give some of the flavor of exactly what is going on in Chinese social media. First we offer two examples, in topic areas without collective action potential, of posts not censored even though they are unambiguously against the state and its leaders. One person wrote, without any hint of collective intent:

> "This is a city government [Yulin City, Shaanxi] that treats life with contempt, this is government officials run amuck, a city government without justice, a city government that delights in that which is vulgar, a place where officials all have mistresses, a city government that is shameless with greed, a government that trades dignity for power, a government without humanity, a government that has no limits on immorality, a government that goes back on its word, a government that treats kindness with ingratitude, a government that cares nothing for posterity. . . "
>
> 这是一个漠视生命的市政府[陕西省榆林市]、一个官员横行的市政府、一个没有公正的市政府，一个低级趣味的市政府，一个包二奶的市政府，一个为钱不要脸的市政府，一个为个权不要人格的市政府，一个没有血性的市政府，一个没有道德低线的市政府，一个出尔反尔的市政府，一个忘恩负义的市政府，一个不要子孙后代的市政府，一个什么怪事都出的市政府，一个什么的市政府，只要你想到的就有…

In critique of the China's One Child Policy, another wrote, without any collective action and without being censored:

"The [government] could promote voluntary birth control, not coercive birth control that deprives people of descendants. People have already been made to suffer for 30 years. This cannot become path dependent, prolonging an ill-devised temporary, emergency measure.... Without any exaggeration, the one child policy is the brutal policy that farmers hated the most. This "necessary evil" is rare in human history, attracting widespread condemnation around the world. It is not something we should be proud of."

可以提倡人民自愿节育，但让人断子绝孙的强制节育，搞30年已是忍辱负重，不能形成路径依赖，将不得已的临时性恶政无限延长.... 可以毫不夸张地讲，计划生育是农民最痛恨的暴政。虽说是"必要的恶"，却是世界少有，遭到世界舆论的广泛遣责，实在不该以此为豪。

Finally in a blog post castigating the CCP for its broken promise of democratic, constitutional government, another wrote, again without any collective action and without being censored:

"I have always thought China's modern history to be full of progress and revolution. At the end of the Qing, advances were seen in all areas, buth after the Wuchang uprising, everything was lost. The Chinese Communist Party made a promise of democratic, constitutional government at the beginning of the war of resistance against Japan. But after 60 years that promise yet to be honored. China today lacks integrity, and accountability starts with Mao. In the 1980s, Deng introduced structural political reforms, but after Tiananmen, all plans were permanently put on hold...intra-party democracy espoused today is just an excuse to perpetuate one party rule. "

我一直将中国的近代史视为一场改良与革命的赛跑，在清末的大赛场上，最终革命跑到了前头，改良的一切设计，在武昌起义枪声响起后成了废纸。

中共的民主宪政承诺，是抗战结束前开出的远期支票，超过一个甲子仍未兑现。当今中国社会缺乏诚信，要从毛泽东开始问责。邓小平在80年代提出的政治体制改革，在"8964"事件后被长期搁置…近年所谓"党主立宪"之说，也是主流学者为维系一党执政地位所做的政治设计。

These posts are neither exceptions nor unusual: We have thousands like these. Negative posts do not accidentally slip through a leaky or imperfect system. The evidence indicates that the censors have no intention of stopping them. Instead, they are focused on removing posts that have collective action potential, regardless of whether or not they cast the Chinese leadership and their policies in a favorable light.

To emphasize this point, we now highlight the obverse condition by giving examples of two posts about events with high collective action potential that support the state but which nevertheless were quickly censored. During the bombings in Fuzhou, the government censored this post, which unambiguously condemns the actions of Qian Mingqi, the

bomber, thus supporting the policies of the government:

"The bombing led not only to the tragedy of his death but the death of many government workers. Even if we can verify what Qian Mingqi said on Weibo that the building demolition caused a great deal of personal damage, we should still condemn his extreme act of retribution... The government has continually put forth measures and laws to protect the interests of citizens in building demolition. And the media has called attention to the plight of those experiencing housing demolition. The rate at which compensation for housing demolition has increased exceeds inflation. In many places, this compensation can change the fate of an entire family. "

爆炸案造成他本人和多名政府工作人员死伤的悲剧，即使钱明奇在微博里所称拆迁造成的个人损失是属实的，我们也应谴责他的极端报复行为....政府在连续出台保护被拆迁者利益的政策法规，媒体也在为公平对待被拆迁者大声疾呼，各地拆迁补偿款的上升速度，大多高于商品房售价上升的速度，在不少地方，补偿款已经足以改变一个家庭的命运。

Another example is the following censored post supporting the state. It accuses dissident Ran Jianxin, whose death in police custody triggered protests in Lichuan, of corruption:

"According to news from the Badong county propaganda department website, when Ran Jianxin was party secretary in Lichuan, he exploited his position for personal gain in land requisition, building demolition, capital construction projects, etc. He accepted bribes, and is suspected of other criminal acts."

湖北省巴东县委宣传部在其官方网站发布新闻通稿称，冉建新在担任利川市都亭办事处党委书记、主任期间，利用职务之便，在征地拆迁、工程发包等事项中为他人谋取利益，收受他人贿赂，涉嫌受贿犯罪。

6 Concluding Remarks

The new data and methods we offer seem to reveal highly detailed information on variation in the interests of the Chinese people, the Chinese censorship program, and the Chinese Government over time and within different issue areas. Using social media to reveal information about those posting is now commonplace, but these results also shed light both on an enormous and secretive government program, as well as on the interests, intentions and goals of the Chinese leadership. The evidence suggests that when the leadership allowed social media to flourish in the country, they also allowed the full range of expression of negative and positive comments about the state, its policies, and its leaders. As a result, government policies sometimes look as bad, and leaders can be as embarrassed, as is often the case with elected politicians in democratic countries, but,

as they seem to recognize, looking bad does not threaten their hold on power so long as they manage to eliminate discussions with collective action potential — where a locus of power and control, other than the government, influences the behaviors of masses of Chinese people. With respect to this type of speech, the Chinese people are individually free but collectively in chains.

Much research could be conducted on the implications of this governmental strategy; as a spur to this research, we offer some initial speculations here. For one, so long as collective action is prevented, social media can be an excellent way to obtain quick and effective measures of the views of the populace about specific public policies and experiences with the many parts of Chinese government and the performance of public officials. As such, this "loosening" up on the constraints on public expression may, at the same time, be an effective governmental tool in learning how to satisfy, and ultimately mollify, the masses. From this perspective, the surprising empirical patterns we discover may well be a theoretically optimal strategy for a regime to use social media to maintain a hold on power. For example, Dimitrov (2008) argues that regimes collapse when its people stop bringing grievances to the state, since it is an indicator that the state is no longer regarded as legitimate. Similarly, Egorov, Guriev and Sonin (2009) argues that dictators with low natural resource endowments allow freer media in order to improve bureaucratic performance. By extension, this suggests that allowing criticism, as we found the Chinese leadership does, may legitimize the state and help the regime maintain power. Indeed, Lorentzen (2012) develops a formal model in which an authoritarian regimes balance media opennes with regime censorship in order to minimize local corruption while maintaining regime stability.

More generally, beyond the findings of this paper, the data collected represents a new way to study China and different dimensions of Chinese politics, as well as facets of comparative politics more broadly. For the study of China, our approach sheds light on authoritarian resilience, center-local relations, sub-national politics, international relations, and Chinese foreign policy. By examining what events are censored at the national level versus a sub-national level, our approach indicates some areas where local governments

can act autonomously. Additionally, by clearly revealing government intent, our approach allows an examination of the differences between the priorities of various sub-national units of government. Because we can analyze social media and censorship in the content of real-world events, this approach is able to reveal insights into China's international relations and foreign policy. For example, do displays of nationalism constrain the government's foreign policy options and activities? Finally, China's censorship apparatus can be thought of as one of the input institutions Nathan (2003) identifies as an important source of authoritarian resilience, and the effectiveness and capabilities of the censorship apparatus may shed light on the CCP's regime institutionalization.

In the context of comparative politics, our work could directly reveal information about state capacity as well as shed light on the durability of authoritarian regimes and regime change. Recent work on the role of internet and social media in the Arab spring (Bellin, 2012; Ada et al., 2012) debate the exact role played by these technologies in organizing collective action and motivating regional diffusion, but consistently highlight the relevance of these technological innovations on the longevity of authoritarian regimes worldwide. Edmond (2012) models how the increase in information sources (e.g., internet, social media) impacts will be bad for a regime unless the regime has economies of scale in controlling information sources. While internet and social media in general have smaller economies of scale, because of how China devolved the bulk of censorship responsibility to internet content providers, the regime maintains large economies of scale in the face of new technologies. China, as a relatively rich and resilient authoritarian regime, with a sophisticated and effective censorship apparatus, is probably being watched closely by autocrats from around the world.

Beyond learning the broad aims of the Chinese censorship program, we seem to have unearthed a valuable source of continuous time information on the interests of the Chinese people and the intentions and goals of the Chinese government. Although we illustrated this with time series in 85 different topic areas, the effort could be expanded to many other areas chosen ex ante or even discovered as online communities form around new subjects over time. Censorship behavior we observe also seems to be predictive of future actions

outside the Internet, is informative even when the traditional media is silent, and likely serve a variety of other scholarly and practical uses in government policy and business relations.

Along the way, we also developed methods of computer-assisted text analysis that we demonstrate work well in the Chinese language and adapted it to this application. These methods would seem to be of use far beyond our specific application. We also conjecture that our data collection procedures, text analysis methods, engineering infrastructure, theories, and overall analytic and empirical strategies might be applicable in other countries.

A Topic Areas

Our stratified sampling design includes the following 85 topic areas chosen from three levels of hypothesized political sensitivity described in Section 3. Although we allow overlap across topic areas, empirically we find almost none.

High: Ai Weiwei, Chen Guangcheng, Fang Binxing, Google and China, Jon Hunstman, Labor strike and Honda, Li Chengpeng, Lichuan protests over the death of Rao Jianxin, Liu Xiaobo, Mass incidents, Mergen, Pornographic websites, Princelings faction, Qian Mingqi, Qian Yunhui, Syria, Taiwan weapons, Unrest in Inner Mongolia, Uyghur protest, Wu Bangguo, Zengcheng protests

Medium: AIDS, Angry Youth, Appreciation and devaluation of CNY against the dollar, Bo Xilai, China's environmental protection agency, Death penalty, Drought in central-southern provinces, Environment and pollution, Fifty Cent Party, Food prices, Food safety, Google and hacking, Henry Kissinger, HIV, Huang Yibo, Immigration policy, Inflation, Japanese earthquake, Kim Jong Il, Kungfu Panda 2, Lawsuit against Baidu for copyright infringement, Lead Acid Batteries and pollution, Libya, Micro-blogs, National Development and Reform Commission, Nuclear Power and China, Nuclear weapons in Iran, Official corruption, One child policy, Osama Bin Laden, Pakistan Weapons, People's Liberation Army, Power prices, Property tax, Rare Earth metals, Second rich generation, Solar power, State Internet Information Office, Su Zizi, Three Gorges Dam, Tibet, U.S.

policy of quantitiative easing, Vietnam and South China Sea, WeiJiabao and legal reform, Xi Jinping, Yao Jiaxin

Low: Chinese investment in Africa, Chinese versions of Groupon, Da Ren Xiu on Dragon TV (Chinese American Idol), DouPo CangQiong (serialized internet novel), Education reform, Health care reform, Indoor smoking ban, Let the Bullets Fly (movie), Li Na (Chinese tennis star), MenRen XinJi (TV drama), New Disney theme park in Shanghai, Peking opera, Sai Er Hao (online game), Social security insurance, Space shuttle Endeavor, Traffic in Beijing, World Cup, Zimbabwe

B Automated Chinese Text Analysis

We begin with methods of automated text analysis developed in Hopkins and King (2010) and now widely used in academia and private industry. This approach enables one to define a set of mutually exclusive and exhaustive categories, to then code a small number of example posts within each category (known as the labeled "training set"), and to infer the proportion of posts within each category in a potentially much larger "test set" without hand coding their category labels. The methodology is colloquially known as "ReadMe," which is the name of open source software program that implements it.

We adapt and extend this method for our purposes in four steps. First, we translate different binary representations of Chinese text to the same unicode representation. Second, we eliminate punctuation and drop characters that do not appear in fewer than 1% or more than 99% of our posts. Since words in Chinese are composed of 1–5 characters, but without any spacing or punctuation to demarcate them, we experimented with methods of automatically "chunking" the characters into estimates of words; however, we found that ReadMe was highly accurate without this complication.

And finally, whereas ReadMe returns the proportion of posts in each category, our quantity of interest in Section 5.3 is the proportion of posts which are censored in each category. We therefore run ReadMe twice, once for the set of censored posts (which we denote C) and once for the set of uncensored posts (which we denote U). For any one

of the mutually exclusive categories, which we denote A, we calculate the proportion censored, $P(C|A)$ via an application of Bayes theorem:

$$P(C|A) = \frac{P(A|C)P(C)}{P(A)} = \frac{P(A|C)P(C)}{P(A|C)P(C) + P(A|U)P(U)}$$

Quantities $P(A|C)$, $P(A|U)$ are estimated by ReadMe whereas $P(C)$ and $P(U)$ are the observed proportions of censored and uncensored posts in the data. Therefore, we can back-out $P(C|A)$. We produce confidence intervals for $P(C|A)$ by simulation: we merely plug in simulations for each of the right side components from their respective posterior distributions.

This procedure requires no translation, machine or otherwise. It does not require methods of individual classification, which are not sufficiently accurate for estimating category proportions. The methodology is considered a "computer-assisted" approach because it amplifies the human intelligence used to create the training set rather than the highly error-prone process of requiring humans to assist the computer in deciding which words lead to which meaning.

Finally, we validate this procedure with many analyses like the following, each in a different subset of our data. First, we train native Chinese speakers to code Chinese language blog posts into a given set of categories. For this illustration, we use 1,000 posts about the labor strikes in 2010, and set aside 100 as the training set. The remaining 900 constituted the test set. The categories were (a) facts supporting employers, (b) facts supporting workers, (c) opinions supporting workers, and (d) opinions supporting employers (or irrelevant). The true proportion of posts censored (given vertically) in each of four categories (given horizontally) in the test set is indicated by four black dots in Figure 9. Using the text and categories from the training set and only the text from the test set, we estimate these proportions using our procedure above. The confidence intervals, represented as simulations from the posterior distribution, are given in set of red dots for each of the categories, in the same figure. Clearly the results are highly accurate, covering the black dot in all four cases.

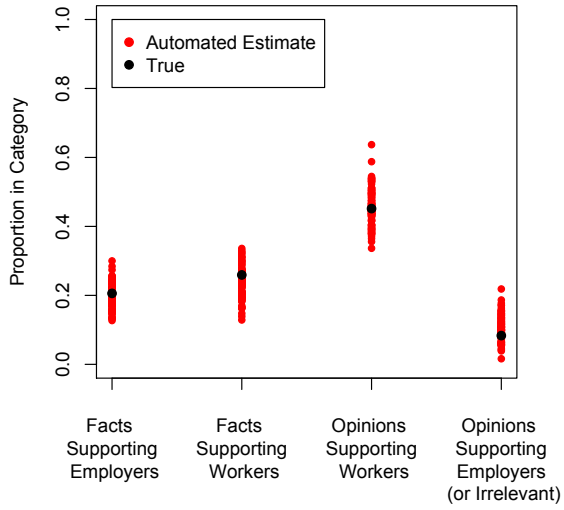


Figure 9: Validation of Automated Text Analysis

References

Ada, Sean, Henry Farrell, Marc Lync, John Sides and Deen Freelon. 2012. "Blogs and Bullets: New Media and Conflict after the Arab Spring.".
URL: *http://www.usip.org/publications/blogs-and-bullets-ii-new-media-and-conflict-after-the-arab-spring*

Ash, Timothy Garton. 2002. *The Polish Revolution: Solidarity*. New Haven: Yale University Press.

Bamman, D., B. O'Connor and N. Smith. 2012. "Censorship and deletion practices in Chinese social media." *First Monday* 17(3-5).

Bellin, Eva. 2012. "Reconsidering the Robustness of Authoritarianism in the Middle East: Lessons from the Arab Spring." *Comparative Politics* 44(2):127–149.

Blecher, Marc. 2002. "Hegemony and Workers' Politics in China." *The China Quarterly* 170:283–303.

Branigan, Tania. 2012. "Chinese politician Bo Xilai's wife suspected of murdering Neil Heywood." *The Guardian* April 10. http://j.mp/K189ce.

Cai, Yongshun. 2002. "Resistance of Chinese Laid-off Workers in the Reform Period." *The China Quarterly* 170:327–344.

Chang, Parris. 1983. *Elite Conflict in the Post-Mao China*. New York: Occasional Papers Reprints.

Charles, David. 1966. The Dismissal of Marshal P'eng Teh-huai. In *China Under Mao: Politics Takes Command*, ed. Roderick MacFarquhar. Cambridge: MIT University Press pp. 20–33.

Chen, Feng. 2000. "Subsistence Crises, Managerial Corruption and Labour Protests in China." *The China Journal* 44:41–63.

Chen, Xi. 2012. *Social Protest and Contentious Authoritarianism in China*. New York: Cambridge University Press.

Chen, Xiaoyan and Peng Hwa Ang. 2011. Internet Police in China: Regulation, Scope and

Myths. In *Online Society in China: Creating, Celebrating, and Instrumentalising the Online Carnival*, ed. David Herold and Peter Marolt. New York: Routledge pp. 40–52.

Crosas, Merce, Justin Grimmer, Gary King, Brandon Stewart and the Consilience Development Team. 2012. "Consilience: Software for Understanding Large Volumes of Unstructured Text.".

Dimitrov, Martin. 2008. "The Resilient Authoritarians." *Current History* 107(705):24–29.

Economy, Elizabeth. 2012. "The Bigger Issues Behind China's Bo Xilai Scandal." *The Atlantic* April 11. http://j.mp/JQBBbv.

Edin, Maria. 2003. "State Capacity and Local agent Control in China: CPP Cadre Management from a Township Perspective." *China Quarterly* 173(March):35–52.

Edmond, Chris. 2012. "Information, Manipulation, Coordination, and Regime Change.". **URL:** *http://www.chrisedmond.net/Edmond%20Information%20Manipulation%202012.pdf*

Egorov, Georgy, Sergei Guriev and Konstantin Sonin. 2009. "Why Resource-poor Dictators Allow Freer Media: A Theory and Evidence from Panel Data." *American Political Science Review* 103(4):645–668.

Esarey, Ashley and Xiao Qiang. 2008. "Political Expression in the Chinese Blogosphere: Below the Radar." *Asian Survey* 48(5):752–772.

Esarey, Ashley and Xiao Qiang. 2011. "Digital Communication and Political Change in China." *International Journal of Communication* 5:298–C319.

Freedom House. 2012. "Freedom of the Press, 2012." www.freedomhouse.org.

Grimmer, Justin and Gary King. 2011. "General purpose computer-assisted clustering and conceptualization." *Proceedings of the National Academy of Sciences* 108(7):2643–2650. http://gking.harvard.edu/files/abs/discov-abs.shtml.

Guo, Gang. 2009. "China's Local Political Budget Cycles." *American Journal of Political Science* 53(3):621–632.

Herold, David. 2011. Human Flesh Search Engine: Carnivalesque Riots as Components of a 'Chinese Democracy'. In *Online Society in China: Creating, Celebrating, and Instrumentalising the Online Carnival*, ed. David Herold and Peter Marolt. New York: Routledge pp. 127–145.

Hinton, Harold. 1955. *The "Unprincipled Dispute" Within Chinese Communist Top Leadership*. Washington, DC: U.S. Information Agency.

Hopkins, Daniel and Gary King. 2010. "Improving Anchoring Vignettes: Designing Surveys to Correct Interpersonal Incomparability." *Public Opinion Quarterly* pp. 1–22. http://gking.harvard.edu/files/abs/implement-abs.shtml.

Huber, Peter J. 1964. "Robust Estimation of a Location Parameter." *Annals of Mathematical Statistics* 35(73).

King, Gary and Langche Zeng. 2001. "Logistic Regression in Rare Events Data." *Political Analysis* 9(2, Spring):137–163. http://gking.harvard.edu/files/abs/0s-abs.shtml.

Kung, James and Shuo Chen. 2011. "The Tragedy of the Nomenklatura: Career Incentives and Political Radicalism during China's Great Leap Famine." *American Political Science Review* 105:27–45.

Kuran, Timur. 1989. "Sparks and prairie fires: A theory of unanticipated political revolution." *Public Choice* 61(1):41–74.

Lee, Ching-Kwan. 2007. *Against the Law: Labor Protests in China's Rustbelt and Sunbelt*. Berkeley, CA: University of California Press.

Lindtner, Silvia and Marcella Szablewicz. 2011. China's Many Internets: Participation

and Digital Game Play Across a Changing Technology Landscape. In *Online Society in China: Creating, Celebrating, and Instrumentalising the Online Carnival*, ed. David Herold and Peter Marolt. New York: Routledge pp. 89–105.

Lohmann, Susanne. 1994. "The Dynamics of Informational Cascades: The Monday Demonstrations in Leipzig, East Germany, 1989-1991." *World Politics* 47(1):42–101.

Lohmann, Susanne. 2002. "Collective Action Cascades: An Informational Rationale for the Power in Numbers." *Journal of Economic Surveys* 14(5):654–684.

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

Lorentzen, Peter. 2012. "Strategic Censorship.".
URL: *http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2101862*

MacFarquhar, Roderick. 1974. *The Origins of the Cultural Revolution Volume 1: Contradictions Among the People 1956-1957*. New York: Columbia University Press.

MacFarquhar, Roderick. 1983. *The Origins of the Cultural Revolution Volume 2: The Great Leap Forward 1958-1960*. New York: Columbia University Press.

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

Marolt, Peter. 2011. Grassroots Agency in a Civil Sphere? Rethinking Internet Control in China. In *Online Society in China: Creating, Celebrating, and Instrumentalising the Online Carnival*, ed. David Herold and Peter Marolt. New York: Routledge pp. 53–68.

Nathan, Andrew. 2003. "Authoritarian Resilience." *Journal of Democracy* 14(1):6–17.

O'Brien, Kevin and Lianjiang Li. 2006. *Rightful Resistance in Rural China*. New York: Cambridge University Press.

Perry, Elizabeth. 2002. *Challenging the Mandate of Heaven: Social Protest and State Power in China*. Armork, NY: M. E. Sharpe.

Perry, Elizabeth. 2008. Permanent Revolution? Continuities and Discontinuities in Chinese Protest. In *Popular Protest in China*, ed. Kevin O'Brien. Cambridge, MA: Harvard University Press pp. 205–216.

Perry, Elizabeth. 2010. Popular Protest: Playing by the Rules. In *China Today, China Tomorrow: Domestic Politics, Economy, and Society*, ed. Joseph Fewsmith. Plymouth, UK: Rowman and Littlefield pp. 11–28.

Przeworski, Adam, Michael e. Alvarez, Jose Antonio Cheibub and Fernando Limongi. 2000. *Democracy and Development: poltical institutions and well-being in the world, 1950-1990*. New York, NY: Cambridge University Press.

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

Ratkiewicz, J., F. Menczer, S. Fortunato, A. Flammini and A. Vespignani. 2010. Traffic in social media II: Modeling bursty popularity. In *Social Computing, 2010 IEEE Second International Conference*. IEEE pp. 393–400.

Reilly, James. 2012. *Strong Society, Smart State: The Rise of Public Opinion in China's Japan Policy*. New York: Columbia University Press.

Rousseeuw, Peter J. and Annick Leroy. 1987. *Robust Regression and Outlier Detection*. New York: Wiley.

Schurmann, Franz. 1966. *Ideology and Organization in Communist China*. Berkeley, CA: University of California Press.

Shih, Victor. 2008. *Factions and Finance in China: Elite Conflict and Inflation*. Cambridge: Cambridge University Press.

Shirk, Susan. 2007. *China: Fragile Superpower: How China's Internal Politics Could Derail Its Peaceful Rise*. New York: Oxford University Press.

Shirk, Susan L. 2011. *Changing Media, Changing China*. New York: Oxford University Press.

Teiwes, Frederick. 1979. *Politics and Purges in China: Retification and the Decline of Party Norms*. Armork, NY: M. E. Sharpe.

Tsai, Kellee. 2007*a*. *Capitalism without Democracy: The Private Sector in Contemporary China*. Ithaca, NY: Cornell University Press.

Tsai, Lily. 2007*b*. *Accountability without Democracy: Solidary Groups and Public Goods Provision in Rural China*. Cambridge: Cambridge University Press.

Whyte, Martin. 2010. *Myth of the Social Volcano: Perceptions of Inequality and Distributive Injustice in Contemporary China*. Stanford, CA: Stanford University Press.

Yang, Guobin. 2009. *The Power of the Internet in China: Citizen Activism Online*. New York: Columbia University Press.

Zhang, Liang, Andrew Nathan, Perry Link and Orville Schell. 2002. *The Tiananmen Papers*. New York: Public Affairs.

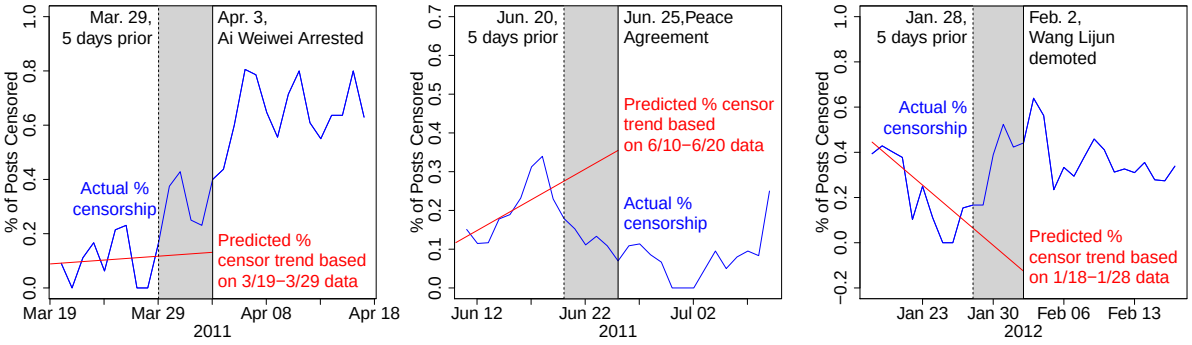
Supplementary Appendix: Prediction as Evidence of Intent

In this section, we offer a final indication that rates and topics of censorship behavior can serve as a measure of the intent of the Chinese leadership. The idea here is that if censorship is a measure of intent to act, then it ought to have some useful predictive value. However, predicting most actions of the Chinese leadership is relatively easy because most of what they do (among that which we observe through the media) are merely responses to exogenous events. The difficult cases for prediction, and those of the most interest from the point of view of understanding China for scholarly and practical policy purposes, are those which are unprovoked, are in some sense voluntary actions that are otherwise unpredictable, and, for our purposes, have collective action potential. We focus on these hard cases here.

We did not design this study or our data collection for predictive purposes, but we can still use it to test our hypothesis. We do this via case-control methodology (King and Zeng, 2001). First, we take all real world events we identified as having collective action potential and remove those easy to predict as a response to exogenous events. This left two events, neither of which could have been predicted at the time they occurred on the basis of information in the traditional news media: the April 3rd, 2011 arrest of Ai Weiwei and the June 25th, 2011 peace agreement with Vietnam regarding disputes in the South China Sea. We analyze these two cases here and show how we could have predicted them from censorship rates. In addition, as we were finalizing this paper in early 2012, the Bo Xilai incident shook China — an event widely viewed as "the biggest scandal to rock China's political class for decades" (Branigan, 2012) and one which "will continue to haunt the next generation of Chinese leaders" (Economy, 2012) — and we happened to still have our monitors running. This meant that we could use this third surprise event as another test of our hypothesis.

Next, we must choose how long in advance censorship behavior could plausibly be used to predict these (otherwise surprise) events. The time interval needs to be long enough so that the censors can do their job, we can detect systematic changes in the percent censored, and so the prediction will have value, but not so long as to make the prediction impossible. We choose five days as fitting these constraints, the exact value of which is of course arbitrary but in our data relatively unimportant. Thus we hypothesize that the Chinese leadership took an (otherwise unobserved) decision to act approximately five days in advance and prepared for it by changing levels of censorship so that they differed from what they would have been otherwise. (Although we do this analysis retrospectively, it was only possible to use as a test because we were checking for censorship rates in real time; going back to check censorship at a later date could induce an artificial relationship solely due to coding rules.)

In Panel (a) of Figure 10, we apply the procedure to the surprise arrest of Ai Weiwei. The vertical axis in this time series plot is the percent of posts censored. The gray area is our five day prediction interval between the unobserved hypothesized decision to arrest Ai Weiwei and the actual arrest. Nothing in the news media we have been able to find suggested that an arrest was imminent. The blue line is actual censorship levels and the red line is a simple linear prediction based only on data greater than five days earlier than the arrest; extrapolating it linearly five days forward gives an estimate of what would have



(a) Ai Weiwei's Arrest

(b) South China Sea Peace Agreement

(c) Wang Lijun Demotion (Bo Xilai affair)

Figure 10: Censorship and Prediction

happened without this hypothesized decision. Then the vertical difference between the red and blue lines on April 3rd is our causal estimate; in this case, the predicted level, if no decision had been made, is at about baseline levels at approximately 10%; in contrast, the actual levels of censorship is more than *twice* as high. To confirm that this result was not due to chance, we conducted a permutation test, using all other 5 day intervals preceding the arrest as placebo tests, and found that the effect in the graph is larger than all the placebo tests.

We then repeat the procedure for the South China Sea peace agreement in Panel (b) of Figure 10. The discovery of oil in the South China Sea led to an ongoing conflict between Beijing and Hanoi, during which rates of censorship soared. According to the media, conflict continued right up until the surprise peace agreement was announced on June 25th. Nothing in the media before that date hinted at a resolution of the conflict. However, rates of censorship unexpectedly plummeted well before that date, clearly presaging the agreement. We also conducted a permutation test here and again found that the effect in the graph is larger than all the placebo tests.

Finally, we turn to the Bo Xilai incident. Bo, the son one of the eight elders of the CCP, was thought to be a front runner for promotion to the Politburo Standing Committee in CPC 18th National Congress in Fall of 2012. However, his political rise met an abrupt end following his top lieutenant, Wang Lijun, seeking asylum at the American consulate in Chengdu on February 6, 2012, four days after Wang was demoted by Bo. After Wang revealed Bo's alleged involvement in homicide of a British national, Bo was removed as Chongqing party chief and suspended from the Politburo. Because of the extraordinary nature of this event in revealing the behaviors and disagreements among the CCP's top leadership, we conducted a special analysis of the otherwise unpredictable event that precipitated this scandal—the demotion of Wang Lijun by Bo Xilai on February 2, 2012. It is thought that Bo demoted Wang when Wang confronted Bo with evidence of his involved in the death of Neil Heywood.

We thus conduct the same analysis for the demotion of Wang Lijun in Panel (c) of Figure 10, and again see a large difference in actual and predicted percent censorship before Wang's demotion. Prior to Wang's dismissal, nothing in the media hinted at the

demotion that would lead to the spectacular downfall of one of China's rising leaders. And for the third of three cases, a permutation test reveals that the effect in the 5 days prior to Wang's demotion is larger than all the placebo tests.

The results in all three cases are strongly confirm our theory, but we conducted this analysis retrospectively, and with only three events, and so further research to validate the ability of censorship to predict events in real time prospectively would certainly be valuable.