Why Autocrats Sometimes Relax Censorship: Signaling Government Responsiveness on Chinese Social Media

Elizabeth Plantan* and Christopher Cairns†

July 12, 2017

Abstract

Despite China's robust censorship capacity, commentary critical of government policies on Chinese social media is ubiquitous. Why would an autocratic regime not fully censor these critiques? We argue that authoritarian leaders periodically relax control to persuade the public that the regime acknowledges citizens' concerns and will address them. This affords state actors a *responsiveness benefit* to weigh alongside other factors, including collective action risk or reputational harm. To illustrate, we use a combination of human- and computer-assisted coding techniques to statistically model censorship of relevant posts on the Chinese microblog *Weibo* during a high-profile air pollution controversy in 2012. We find two distinct trends in censorship around a crisis event during which the state largely relaxes control. After the crisis, leaders adjust to allow some limited critiques, while blocking directly disparaging remarks. This suggests that the state changes censorship in order to signal responsiveness to citizens' legitimate concerns over governance.

Keywords

Authoritarianism; information control; censorship; social media; China; air pollution

*PhD Candidate, Cornell University. Email: cmc467@cornell.edu

†Research Analyst, China Studies, Center for Naval Analyses. Email: cmc467@cornell.edu

Introduction

Despite robust censorship capabilities, sporadically open public debate in the Chinese blogosphere persists. Internet users can discuss their opinions about such sensitive topics as the Diaoyu/Senkaku islands conflict (Second Author & Co-Author, 2016) or food safety concerns (Yang, 2013), with comments ranging from implicit or guarded criticism of government policies (Esarey & Xiao, 2008) to outright criticism of the regime itself. Why would an autocratic regime with the capacity to control public information flows not fully censor these critiques?

Studies of information control in non-democratic regimes address this question by focusing on how authoritarians oscillate between liberalization and repression. On the one hand, an authoritarian leader needs information about their true level of support, on the other, if regime-disparaging information were to spread freely, it could fuel coordination among opponents to overthrow the regime. This problem, coined the "dictator's dilemma" (Wintrobe, 1998), is a central tension of authoritarian rule. To solve it, authoritarian leaders have to balance between 1) allowing enough genuine dissent to gather information about dissatisfaction and 2) maintaining control through repression or co-optation. The extant literature focuses on how authoritarians solve this dilemma by keeping credible information flowing vertically from citizens to the state, while suppressing or co-opting potentially harmful horizontal information transfers from citizen-to-citizen.[1] But what if an authoritarian leader wanted to use these vertical information channels to send a credible signal back to citizens? What if the dictator not only wants to gather information from the public by relaxing control, but also wants to convey information to the public by doing so?

In this paper, we argue that authoritarian leaders can (temporarily) relax control over dissent to persuade citizens that the regime acknowledges their concerns as legitimate and will be responsive to their demands in the future. This strategy affords the regime a *responsiveness benefit* that state actors weigh against other concerns, such as a collective action risk or harm to regime legitimacy, to find an optimal balance between liberalization and repression. Our concept differs from previous explanations of why non-democratic

leaders relax control, such as the idea of providing a "safety valve" for the public to vent grievances (Hassid, 2012; MacKinnon, 2008) or managing dissent through a "controlled burn" (Lorentzen, in press), because it not only conveys public concern to the state, but also allows the state to send a credible signal to the public that its legitimate concerns will be addressed. We illustrate this concept with a case study of online social media discussion of air pollution in China.

This case is ideal for illustrating our argument for several reasons. First, China is a world leader among non-democracies with respect to information control over media. Observing relaxed censorship of online dissent in China is no accident since the state has the proven capacity to use nuanced forms of censorship to filter or block out undesired information (King, Pan, & Roberts, 2013). Second, choosing social media (Sina *Weibo*, China's version of Twitter), instead of traditional media, allows for an examination of state-society interaction in a space of "counter-hegemony" (Yang, 2013). Although censors manage online discussion and government-paid commenters (*wumaodang*) abound, the majority of content on social media is still citizen-generated and spontaneous, especially during crises or scandals. It also differs from the formalized "letters and visits" (*xinfang*) system because the information stream is public. It not only facilitates horizontal information transfer between Internet users, but also allows the state to transfer information to the Internet-using public through its management of that space, generating "common knowledge" (Kuran, 1995) between citizens about the state's intent. Finally, we choose a controversy about air pollution in 2012 because it raises the visibility of censorship, allowing citizens to perceive the government's actions (M. Roberts, 2015) with respect to managing online discussion of the conspicuous problem of air pollution.

Although relaxing censorship allows the government to signal responsiveness to the public, the regime cannot completely relax control over dissent at all times due to the constraints of the "dictator's dilemma." Instead, we argue that they make nuanced decisions about which information to allow or block, sending a signal to the public about responsiveness while containing the most volatile public sentiment. To illustrate this, we consider three distinct framings (or "sentiments") of the problem with different levels of

perceived risk to regime stability and show how they are censored differently through a combination of hand-coded and computer-assisted content analysis of Weibo posts, as well as statistical modeling of sentiment trends and their rates of deletion. We find that censorship of air pollution can be separated into two distinct time periods around a crisis event during which censorship was largely relaxed and then adjusted to match the regime's new priorities. This suggests that when autocrats choose *not* to censor, their primary interest is not in incoming information flows from society, but in outward communication: signaling their responsiveness to growing popular demands through the release of netizens' own voices.

Relevant Literature

The quandary that authoritarians face between gathering information and maintaining control stems from the concept of the "dictator's dilemma" (Wintrobe, 1998). As a dictator becomes more powerful and repressive, it becomes harder for him to obtain information about his true level of support because citizens (and even elites) are afraid of looking disloyal. These actors engage in "preference falsification," participating in ritualistic shows of support for the regime to hide their true feelings and protect themselves from the leader's wrath (Kuran, 1995). To address this dilemma, leaders must balance between allowing enough genuine dissent to gather information about dissatisfaction and maintaining control through repression or co-optation. Recent literature on the threat of elite discontent has considered how institutions, such as legislatures (Gandhi, 2008; Gandhi & Przeworski, 2006), parties (Brownlee, 2007; Magaloni, 2006, 2008), and elections (Blaydes, 2011; Gandhi & Lust-Okar, 2009), co-opt potential opposition and publicly signal the leader's commitment to share power (Boix & Svolik 2013; Svolik 2012). Others have focused on how authoritarians gather credible information from citizens, such as through public opinion polls, formal citizen complaints (Dimitrov 2014a; 2014b; 2015; Wang and Peng, 2015), or even limited public protest (Lorentzen 2013). However, these information-gathering methods can backfire, since they could in-

crease citizens' "common knowledge" (Kuran 1995) of each other's discontent and fuel regime-toppling "information cascades" (Lohmann, 1994).

Recent studies of authoritarian information control focus on the potential for either regime stability or collapse, often over-looking the more complex (and less threatening) state-society interactions that occur through the everyday management of information in authoritarian regimes. Formal modelers that focus on the tradeoff between the need for information and the risk of spreading popular discontent (Egorov, Guriev, & Sonin, 2009; Gehlbach & Sonin, 2014; Whitten-Woodring & James, 2012) typically only give the government a dichotomous option between repression and liberalization. These studies tend to focus on traditional media, but even those that consider online content and social media (Little, 2016; Reuter & Szakonyi, 2015) focus on the state's dichotomous options for limiting the Internet's potential for facilitating collective action that could lead to regime collapse. Within the study of Chinese information control, scholars focus on the government's impressive capacity for censorship of information (MacKinnon, 2008; Morozov, 2011; Yang, 2009; Zheng, 2007), motivated by fears of collective action (King, Pan, & Roberts, 2013; 2014), the desire to preserve leaders' reputations (Esarey, 2013), or simply because censorship can go undetected, thus lowering its cost (M. Roberts, 2014). Most of these studies conceive of information flowing one way vertically from the public to the dictator and horizontally from citizen-to-citizen (Lorentzen, in press). However, just as limited power-sharing institutions provide a "public observably signal of the dictator's commitment" (Svolik, 2012, p. 8), so can the dictator's management of publicly available information flows signal commitment or intent to the public.

Other studies of non-censorship address the idea that the state might gain from releasing control over information (Lorentzen 2014), but the focus remains on the vertical flow of information from citizens to the state. Hassid (2012) claims that allowing the airing of public grievances acts as a "safety valve" for citizens to vent their feelings. Lorentzen (in press) argues that the more appropriate metaphor is a "controlled burn" in which the fire of public opinion is allowed to break out in order to burn off in a contained area (p. 9). In both metaphors, the state permits this airing of discontent at some risk to itself

(since safety valves can explode and controlled burns can turn into wildfires), in order to diffuse worse problems later on. The flow of information is still vertical from citizens to the state, or horizontal between citizens. But public information-gathering channels, like social media, are two-way streets between state and society. Not only can the state use these public channels to gather information, but citizens viewing them can draw inferences about the state's commitment to meet citizen demands based on the level of information control. Observing non-censorship or relaxed censorship could be a strategic signal of responsiveness from leaders to citizens. We develop this idea in the following theoretical section.

Information Management under Authoritarianism: A Four-variable Framework

We argue that authoritarian regimes with high, centralized capacity to control information are subject to four rational cost-benefit considerations that strongly shape decisions to block or release information: *responsiveness benefit*, *image harm*, *visible censorship cost*, and *collective action risk*. While the latter three factors are rooted in existing work, the concept of *responsiveness benefit* is new. By allowing citizens to openly discuss a sensitive policy issue, the government implicitly signals acknowledgment of the problem and its intent to address it, gaining a degree of *responsiveness benefit*. Conversely, if citizens express concern about a problem but observe swift control, they might infer that leaders are either not being responsive to citizens' legitimate demands or that the public airing of grievances about the problem is politically unacceptable. The logic of *responsiveness benefit* presumes that citizens both observe when censorship occurs and interpret non-censorship as a signal intended by leaders not only to appear responsive, but also strong and capable of immediately addressing public concern. By allowing discussion, authoritarian leaders recognize that citizens will generate collective perceptions of leaders' responsibility to fix the problem, including some amount of criticism. But leaders may hope that the majority of citizens will perceive even limited openness as a

signal that real reform is around the corner (as in the "Democracy Wall" movement of 1978-79 which presaged major 1980s reforms in China). Thus, the idea of *responsiveness benefit* suggests an *interactive* form of information management in which the state acknowledges popular grievances and communicates its intent to address them.

This is not to suggest that an authoritarian leader will relax censorship at all times, however, since leaders still weigh the risks and benefits of this approach to the problem of information versus control. At the same time that leaders consider *responsiveness benefit*, they must also balance between the other factors: *image harm*, *visible censorship cost*, and *collective action risk*. The concept of *image harm* is defined as the probability that a majority of citizens will interpret non-censorship not as positive acknowledgment of a problem, but as evidence of a weak, ineffective or divided central leadership (as in Esarey, 2013). Even if mobilization does not occur immediately, inferring state weakness or ineptitude (and generating shared knowledge of this fact through open discussion) increases the potential for anti-regime activity later on. In this way, *responsiveness benefit* and *image harm* can be viewed as part of the same calculation, where one denotes citizens' shared perceptions of regime strength, and the other, weakness. We term this the state's *credibility payoff*, according to Equation 1:

$$\textit{credibility payoff} = \textit{responsiveness benefit} - \textit{image harm} \qquad (1)$$

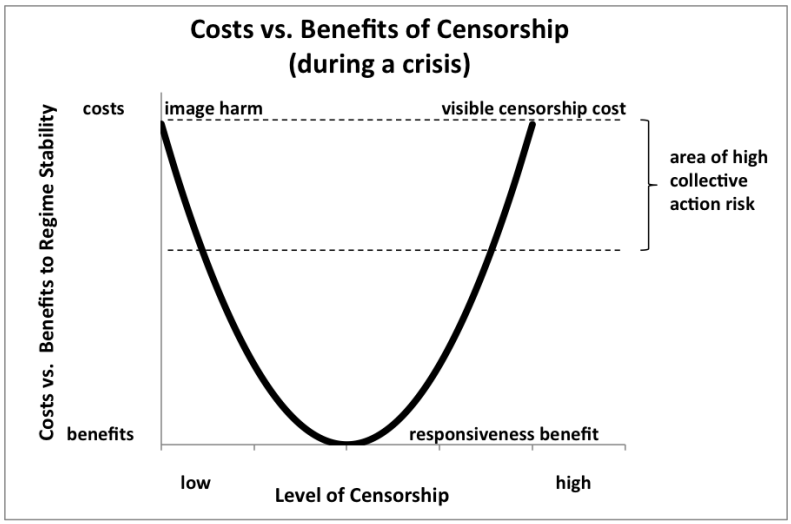
The third factor, *visible censorship cost*, addresses citizen awareness of being censored. Studies (M. Roberts 2014; 2015) show that when censorship is sufficiently invisible – citizens are unaware or unsure that they are being censored and unable to find any relevant information – they give up. But if individuals are aware of censorship, they may be *more* motivated to post sensitive information. Therefore, highly visible censorship during high-profile political events can backfire, depending on the public's pre-existing awareness of the problem and the availability of alternative information sources. Thus, if leaders think that censorship will be too visible and cause harm to regime legitimacy, they may refrain from cracking down.

The fourth and final factor, *collective action risk*, has received substantial empirical support (King, Pan, & Roberts, 2013; 2014). Social media posts during so-called "topic

bursts" (surges in online discussion on a specific topic) that relate to real-world collective actions such as street protests are more likely to be censored. However, although high censorship of posts at a highly volatile time might help diffuse collective action risk, it could also provoke it if censorship is visible. If citizens can perceive censorship, then they might decide that a topic is off limits and quiet their dissent, or they could assume that the government does not intend to be responsive to their demands and may then have greater incentives to take to the streets (Meng, Pan, Hobbs & Roberts, 2017). Furthermore, if an authoritarian regime chooses *not* to censor during a "topic burst" of online commentary, it suggests that *collective action risk* alone is not the primary consideration.

During crisis events that cause a "topic burst" in online dissent, the state may have to re-consider its censorship actions with respect to *responsiveness benefit*, *image harm*, *visible censorship cost*, and *collective action risk*. Crises can be caused by exogenous shocks (such as natural disasters) or by the government's own actions that may have an unintended impact on public support (such as an international incident) at the national level. This type of national-level consciousness is necessary for the state to receive a *responsiveness benefit*, since the state benefits from its response to a problem that is highly visible and important to large numbers of citizens. The costs and benefits to the state of its chosen information management strategy during a crisis event are U-shaped, as depicted in Figure 1.

Figure 1: Costs vs. Benefits of Censorship



If the regime did not censor at all, then critical information would fly freely and the regime would be perceived as weak, which could increase *image harm* and contribute to *collective action risk* in the long term. If there is some censorship, then state actors could shape the conversation and show that they are listening and responding to public demands, receiving a *responsiveness benefit*. If the regime censors the issue completely, then the public might perceive that the regime is not being responsive (thus increasing *visible censorship cost*) and potentially increase the incentives for citizens to take their grievances to the street (increasing *collective action risk*). Thus, the state must balance between all four of these factors when choosing how to respond.

To observe a crisis-motivated shift in censorship strategy as the state rebalances between these four factors, we categorize three main stages. The first is the state's *Business-as-usual* phase during which government censors follow standard operating procedures. In China, in-house censors at Internet companies have lists of banned keywords to guide deletion of online content for topics that the state deems "off limits." While censors do their best to repress a banned topic in its early stages, netizens can sometimes evade these measures by altering the words they use. Furthermore, public pressure sometimes becomes so strong that the state is prompted to reconsider its approach. Rather than doubling down on censorship, our argument is that sometimes leaders reach a turning point – a *signal shift* – followed by an *Adaptive phase* where they re-assess the tolerable limits of legitimate citizen criticism, while more aggressively filtering destabilizing comments. We expect to see these dynamics as the state balances among the factors in our theoretical framework, which is explored further in our context-specific case study.

Case Selection: Censorship of Air Pollution in China

To illustrate our framework, we examine Chinese censorship of air pollution-related commentary on the social media platform Weibo. The discussion of a sensitive and highly-visible problem on social media is a "most likely" case for illustrating a shift in information management under authoritarianism. First, China has the documented ability to control information within its borders with "military-like precision" (King, Pan, & Roberts, 2013,

p. 1), rapidly and effectively responding to breaking incidents (Bamman, O'Connor, & Smith 2012; King, Pan, & Roberts, 2013; 2014; Zhu, Phipps, Prigden, Crandall, & Wallach, 2013) according to centralized top-down directives (Second Author, 2016). Second, the seriousness and scale of environmental issues have aroused public concern, yet discussion in this area remains relatively open (Hildebrant & Turner, 2009; Ho, 2001; Ho & Edmonds, 2008; Yang & Calhoun, 2007). Since environmental activism enjoys a privileged space in Chinese domestic politics, fluctuating from repression to toleration and even encouragement, it is a likely topic to observe variation in information management. Air pollution, in particular, is an environmental problem that has arguably attracted the most attention in China among the urban, educated, Internet-using public. We examine discussion of this highly visible issue on Weibo, which was at its peak as a lively public forum during the period under study. This platform is ideal for illustrating the concept of *responsiveness benefit* since the temporary relaxing of censorship would be obvious in such a public digital space.

To narrow further, we focus on a controversy over the release of air quality monitoring data that became a political flashpoint between the U.S. Embassy in Beijing and the Chinese government in 2012. Although daily Air Quality Index (AQI) data has been available in many Chinese cities since the early 2000s, the issue became more contentious when the U.S. Embassy in Beijing began including measurements of PM 2.5 (particulate matter of 2.5 micrometers in diameter or less) in this data in 2008, making it more fine-grained than the official data that only included the larger PM 10 (Chan & Yao, 2008).[2] In 2012, on World Environment Day (June 5), after years of private complaints about the U.S. Embassy's data release, Vice-Minister of Environmental Protection Wu Xiaoqing publicly accused the U.S. of violating China's sovereignty (Bradsher, 2012). On the morning of June 6, several newspapers reported Wu's remarks and set off a Weibo firestorm of negative reactions. Many commenters lambasted the Chinese government for not tackling the problem of air pollution head on or releasing its own data, thus forcing the U.S. Embassy to undertake what should have been the government's responsibility.

Weibo commentary about Wu's remarks continued to simmer for several days after

June 5 and 6. Just as it began to wane, on June 12 Vice Foreign Affairs Minister Cui Tiankai re-ignited the controversy by stating that foreign embassies should not be expected to improve China's air quality, but rather, the Chinese people should be held accountable for improving the situation (Henochowicz 2012). The next day (June 13), netizen responses to Cui were even more mocking than before, with bloggers accusing him of attempting to divert blame away from what many viewed as a government cover-up.

These dates (June 6 and 13) represent responses to two very high-profile public comments by top Chinese officials and also book-end the period with the highest Weibo activity related to air pollution for the entire year. June 13 is the peak of public mobilization, when we would expect the government's *responsiveness benefit* from non-censorship to be at its height, after which we expect the government to adapt to the crisis and reassert control over hostile speech, while accommodating perceived non-threatening issue framings. Therefore, we argue that the period before June 6, 2012 is the state's *Business as usual* phase, the period between June 6 and 13 represents its *signal shift* as it relaxes censorship to receive a *responsiveness benefit*, and that the period after June 13 represents the state's *adaptive phase* as it adjusts its censorship strategy. The following empirical analysis provides quantitative evidence for the shift that we have qualitatively described here.

Data and Method

To analyze censorship on Weibo, we rely on the WeiboScope dataset collected by Fu, Chan, and Chau (2013). The dataset consists of posts from over 38,000 Weibo users with verified identities as public figures and more than 10,000 followers as of January 2012. These users have greater resonance with the broader Weibo community and contribute to the volatility of online commentary. Each row in the dataset consists of one social media post plus associated meta-data. We analyze the post text and count embedded reposts as part of the text.

Our main dependent variable is the censorship rate, defined as the number of posts

recorded as censored in the WeiboScope data divided by total topic-relevant posts per day. The WeiboScope dataset uses a program to measure censorship by checking for deleted posts every 24 hours. However, some fraction of posts could be deleted prior to the program taking its daily record, which means that the actual censorship rate may be much higher than the dataset suggests. To address this problem we use a mathematical correction based on prior work (Second Author & Co-author, 2016) to estimate the "true" censorship rate.[3]

Censorship Predictions by Sentiment Category

To examine nuanced differences in censorship, we assign posts into categories based on a close reading of the Weibo data. Discussion of air pollution on Weibo focuses on three main sentiments or "frames" of the issue that vary in their level of risk to regime stability: 1) *political* criticism; 2) concerns about *physical harm*; and 3) *scientific* information. The fluctuations of these categories with respect to the daily censorship rate are observable implications of how the regime balances between *credibility payoff* (the difference between *image harm* and *responsiveness benefit*), *visible censorship cost*, and *collective action risk*. We score each independent variable according to how authoritarian leaders perceive its risk or benefit. Overall, the *political* category poses the most risk, while the *scientific* category has the most potential benefit. The *physical harm* category – with discussions centering around the human health harms of air pollution – could be both beneficial (for example, the government releasing PM 2.5 data to raise health awareness) or costly (for example, pointing out WHO standards that Beijing is not following). For this reason, we have coded it as a medium risk. These categories are laid out in Table 1.

Table 1: Regime Cost/Benefit by Sentiment Category

Category	Level of Risk or Benefit
Political	High Risk, Low Benefit
Physical Harm	Medium Risk, Medium Benefit
Scientific	Low Risk, High Benefit

We also consider how the crisis event in June (our *signal shift*) may affect the regime's calculus. During the *business as usual (BAU) phase*, *visible censorship cost* is relatively low, since the public has no external event with which to observe an increase in censorship from the regime.[4] If the regime can get away with censorship undetected, then it can minimize *image harm*. However, by not addressing the air pollution issue, it cannot receive any *responsiveness benefit*. At this time, *political* comments about air pollution are considered high risk and censored accordingly, while comments about *physical harm* are relatively neutral (medium risk), and *scientific* comments are low risk. The regime would censor at its *Business as usual* rate without a sense of urgency (censoring a few days after comments are posted). During the crisis event (at the moment of a *signal shift*), when *visible censorship cost* is at its highest, the regime immediately drops its censorship rate to allow public opinion to flourish. This is how it receives its *responsiveness benefit*, although it must also consider the *image harm* and *collective action risk* from doing so. After the crisis (in the *Adaptive phase*), government censors adjust based on different framings of the issue. Once the scandal has died down, visible censorship cost is lower, allowing the regime to censor more. This allows it to take back control over potential *image harm* and *collective action risk*. During this latter phase, *political* commentary is very high risk, *physical harm* remains neutral, and *scientific* comments are very low risk and even beneficial to the regime. By allowing *scientific* framings of the air pollution issue to continue, the regime is showing its intent to address the PM2.5 monitoring issue, thus securing the *responsiveness benefit* gained during the *signal shift*. The speed of censorship also adapts to the sensitivity of the topic, increasing in response to the most negative and critical comments. These shifts in censorship are mapped in Table 2.

Table 2: Predicted Censorship According to the Four-Variable Framework

Collective Action Risk	*Credibility Payoff*	*Visible Censorship Cost* **High**	**Low**
Low	**Positive**	Low Censorship	Low-Medium Censorship
	Negative	Medium-High Censorship	High Censorship
High	**Positive**	Low-Medium Censorship	Medium-High Censorship
	Negative	High Censorship	Very High Censorship

Note: Recall that *credibility payoff* can be either positive or negative because it is the difference of *responsiveness benefit* and *image harm*.

Table 2 predicts relative levels of high or low censorship that result from applying the four-variable framework separately to each sentiment category, depending on that category's unique level of risk. It also contextualizes the state's changing censorship strategy before and after the *signal shift*, during the *Business-as-usual* and *Adaptive* phases. For example, during the *Business-as-usual* phase, *visible censorship cost* would be lower, allowing the state to censor more, whether *credibility payoff* from a certain sentiment category is positive or negative. However, when *visible censorship cost* is higher during the *signal shift* or *Adaptive phase*, the regime may censor sentiments with a positive *credibility payoff* less (such as *scientific* commentary) and those with a negative *credibility payoff* (as in *political* commentary) slightly more. This is all conditioned by high or low levels of *collective action risk*, which can heighten or lessen censorship in the whole system.

Coding Methods and Measures

To filter out only pollution-relevant discussion in the WeiboScope dataset, we created our sample from posts containing one or more of the following keywords: "air pollution" (*kongqi wuran* or *daqi wuran*), "air quality" (*kongqi zhiliang* or *daqi zhiliang*), "smog" (*wumai*), "haze" (*huimai* or *huiwu*), and "PM 2.5" (in Latin characters). This left 71,088 relevant posts for all of 2012 to use as our coding sample. Given the high volume of posts, we used a combination of human- and computer-assisted coding techniques that included several pre-coding stages.[5]

For the *political* category, we included three measures. In light of the U.S. Embassy's

release of PM2.5 data, Chinese bloggers often compared the air quality situation in their own country to other countries or to the international community, a phenomenon we term "Domestic vis-à-vis Foreign." While codings of this measure included both pro- and anti-state commentary, we found that a large majority of such comments reflect poorly on Beijing's handling of the problem. A second category captured whether posts assigned any responsibility (or even blame) to the Chinese government either for having allowed air pollution to worsen, or for not doing enough to clean it up. We labeled comments in this category with negative valence as "Anti-Government". Our third and final *political* measure was the keyword "U.S. Embassy" (*shiguan*) itself, which we found to proxy well for politically critical speech on the issue of air pollution in 2012.[6] Combined, these three separate measures comprise our measure of *political* speech.

For the *physical harm* category, we included a hand-coded measure of whether air pollution-related comments framed the issue as a threat to human health (labeled "Health") and a keyword count of the Chinese word for health ("*jiankang*"). Third, our *scientific* category contained two measures. The first, "AQI Monitoring", is a human-coded measure of whether a post primarily contained air quality monitoring statistics. We also measure the keyword "PM 2.5". Although the term appeared in a variety of contexts, we included it in our *scientific* category because it refers to a scientific standard for measuring air pollution, and thus connotes scientific legitimacy even when embedded in more politically sensitive speech.

Finally, we include two additional measures as controls. We measure the presence of "News" by counting all posts containing a left bracket ("[") which nearly always signifies the beginning of a news story link. Spikes in pollution-relevant media reporting may be related to the prevalence of certain sentiment categories and to the censorship rate. As an additional control we include actual Air Quality Index (AQI) data from the Beijing U.S. Embassy's monitoring station in 2012 ("real-time AQI") to condition all of our results on real-world pollution fluctuations.

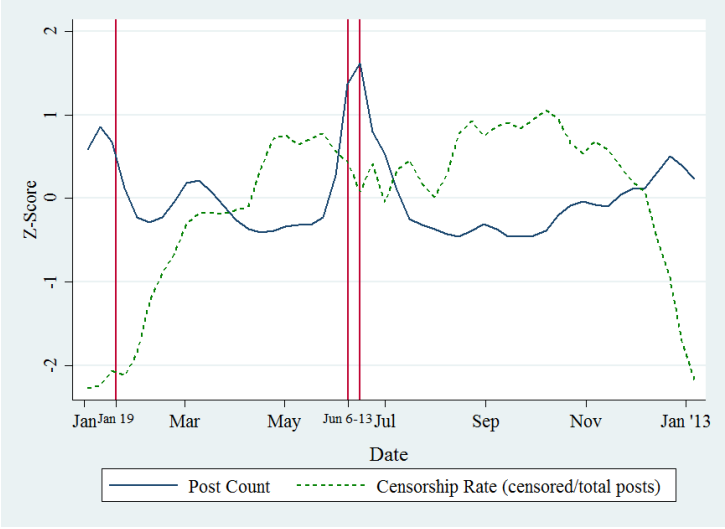
To code these measures, we draw on a random sample of 500 posts across the whole year then sub-sample 150 posts from the two dates that mark our *signal shift* (June 6

and June 13). This allows us to compare the period before the *signal shift* (the *Business as usual* phase) to the period after (the *Adaptive* phase) and observe how Chinese leaders shift censorship strategy. We also include data on year-long trends in censorship across all sentiment categories. Since hand-coding over 71,000 posts is infeasible, we use a computer assisted text analysis (CATA) algorithm called *ReadMe* (Hopkins & King, 2010) to estimate the proportions for the entire year.[7]

Results

First, we view our sample of 71,088 relevant posts drawn from the WeiboScope dataset compared to censorship. Figure 2 presents the standard deviations for both the censorship rate and relevant posts, showing which dates are above or below the average daily censorship rate or post count for the year. Relevant posts peak on several dates throughout the year. The first peak correlates with the worst real-time air pollution measurement on January 19.[8] The second "volume burst" happens in June around our key *signal shift* dates of June 6 and June 13.[9]

Figure 2: Relevant Posts and Censorship Rate for 2012
(Moving Average)



When post volumes surge in June, the censorship rate falls relative to its daily long-term average for 2012. Similarly, during the high-pollution winter with a slightly higher

relevant post count, censorship is below its mean. One reason for this is that when air pollution is visibly bad, harsh censorship could backfire, since *visible censorship cost* would be high. Similarly, when posts about air pollution surge in June, not only is *visible censorship cost* high, but national attention to the issue means that the state could receive its maximum *responsiveness benefit*. This suggests that the June events serve as a *signal shift* before the regime adjusts censorship.

Next, we consider summary statistics of our keywords and *ReadMe*-estimated sentiment measures in Table 3. During the *Business as usual* phase the PM 2.5 keyword was the most widespread, with AQI monitoring data and health-related commentary also prominent. In addition, the proportion of news stories was substantial. *Political* criticisms (Domestic vis-a-vis Foreign and Anti-Government speech) were also high when compared with the latter half of the year during the *Adaptive* phase. Finally, the censorship rate, though not low in absolute terms, was lower (.49) than the year-long average (.57).

With this period as a baseline, the June 6 and June 13 measures are put into context. On June 6, measures for most categories are relatively low, although health-related commentary, the PM 2.5 keyword, and news are similar to the *Business as usual* phase. The censorship rate is quite high (.71) when compared to the year-long average (.57). On June 13, this changes. The *political* measures (Domestic vis-a-vis Foreign, Anti-Government, and the U.S. Embassy keyword) are very high. Concerns about health are also high and AQI monitoring rebounds to its value in the *Business as usual* phase. Discussions of PM 2.5 are minimal and news is a very small proportion of posts. Remarkably, the censorship rate for June 13 is unusually low. This suggests that despite the potential harm of *political* speech, government censors chose to relax control during the highest "topic surge" on June 13.

During the *Adaptive* phase, our *political* measures showed marked declines compared with earlier in the year, particularly Anti-Government (.34 to .21). In the *scientific* category AQI Monitoring increasingly dominated the topic blend, but mentions of PM 2.5 also remained moderate. News was the next highest proportion, slightly surpassing

Table 3: Sentiment Categories by Estimated Measures

Sentiment category	Measure	BAU phase (Jan 2-Jun 5)	Signal shift (Jun 6)	Signal shift (Jun 13)	Adaptive phase (Jun 14-Dec 30)
Political (high risk)	Domestic-vis-a-vis-foreign	.22	.18	.83	.16
	Anti-government	.34	.23	.84	.21
	U.S. Embassy (keyword)	.04	.25	.69	.02
Physical harm (medium risk)	Health	.28	.24	.66	.22
	"Jiankang" (keyword)	.08	.03	.02	.07
Scientific (low risk)	AQI monitoring	.29	.11	.25	.42
	PM 2.5 (keyword)	.38	.32	.12	.25
Additional measures	News ("[")	.33	.28	.09	.37
	Real-time AQI (U.S. Embassy)	97	143	70	87
	Censorship rate	.49	.71	.30	.64
	# of posts	181	1460	2363	164

the level in the *Business as usual* phase. In the *physical harm* category, discussions of health-related concerns are at a moderate level. Finally, the censorship rate showed a substantial increase compared to both June 13 and before. Overall, these proportions suggest that leaders allowed less threatening sentiments to become increasingly prevalent after June 13, while more threatening *political* speech was increasingly restricted.

Overall, the summary statistics show a general difference in category proportions and censorship between the *Business as usual* and *Adaptive* phases and a dip in censorship relative to *political* sentiment during the June events. Next, we model the relationship between these proportions and the censorship rate.

Modeling the Sentiment Categories' Relation to Censorship

To show the shift in censorship after June, we compare regression models for the *Business-as-usual* phase to the *Adaptive* phase. We expect the directions of significant effects to resonate with Tables 1 and 2: *political* measures should positively correlate with censorship, *physical harm* measures should show weak or no relation, and *scientific* measures should be negatively correlated. During the *Adaptive* phase, the relationships between censorship and the *political* and *scientific* measures should be in the same direction as before, but intensify. Because we have time series data, we use Generalized Linear Model (GLM) regression and assume that the censorship rate has a binomial distribution and that the model takes a logistic form. We address autocorrelation by using Newey-West standard errors.[10] Table 4 shows results during the *Business as usual* phase.

Table 4 presents four model specifications and displays lags zero and one.[11] Model I consists only of the key sentiment measures for *political* and *scientific*. The *physical harm* measures are absent from the baseline model because we found that this was our noisiest measure, possibly because health-related concerns tended to be captured alongside other concerns like foreign comparisons. For this reason, we add the *physical harm* measures only in Model IV. Next, Model II adds News, and Model III further adds real-time AQI. For all models, lag one of the censorship rate is positive, significant and large. This autoregressive characteristic, where censorship on Day 1 predicts censorship on Day 2,

Table 4: Censorship During the BAU Phase (Average Marginal Effects)

DV: Cens. Rate	Model I	Model II	Model III	Model IV
L.Cens. Rate	0.269***	0.275***	0.281***	0.288***
Political				
Dom. v. For.	-0.005	-0.005	-0.010	-0.095
L.Dom. v. For.	0.017	0.020	0.012	0.012
Anti-Govt	-0.004	-0.002	0.013	0.024
L.Anti-Govt	0.140***	0.158***	0.150***	0.131***
U.S. Embassy	0.038	0.069	0.075	0.073
L.U.S. Embassy	0.202	0.196	0.276**	0.239
Scientific				
AQI Monitoring	0.091	0.105*	0.089	0.090
L.AQI Monitoring	-0.164**	-0.198***	-0.170**	-0.171**
PM 2.5	-0.050	-0.060	-0.034	-0.016
L.PM 2.5	-0.150**	-0.173**	-0.153*	-0.155*
Controls				
News		-0.043	-0.031	-0.064
L.News		0.126*	0.094	0.068
Real-time AQI			0.099*	0.097*
L.Real-time AQI			-0.103*	-0.123**
Physical Harm				
Health				0.072
L.Health				0.005
Jiankang				0.379*
L.Jiankang				0.066

$$^{*}\ p < 0.1\ ^{**}\ p < .05\ ^{***}\ p < .01 \qquad N = 151$$

is expected. After a breaking incident, censors delete the majority of targeted content shortly thereafter, but the censorship rate typically remains high for a few days.

For the explanatory variables, Anti-Government lag one is positive and significant in all models. This suggests that government censors are particularly attuned to this *political* speech during the *Business as usual* phase. The other two measures that stand out are *scientific*. Both AQI Monitoring and mentions of PM 2.5 lag one are negatively correlated with censorship across all four models. This suggest three points: first, that the censors differentiate between the scientific, "objective" information captured by these measures versus most other forms of Weibo content; second, that even controlling for PM 2.5

mentions appearing as part of AQI Monitoring, the PM 2.5 keyword is censored less; and third, that AQI Monitoring predicts reduced censorship despite its frequent co-occurrence with keywords that predict the opposite, suggesting that censors may distinguish between air monitoring reports from Chinese sources versus the U.S. Embassy.

For the other *political* measures, U.S. Embassy lag one is only positive, large, and significant in Model III, once real-time AQI is added but before Health is added in Model IV. A measure for *physical harm* is only significant when "Jiankang" is added in Model IV. This may be because both the "U.S. Embassy" and "Jiankang" keywords were closely related. Many were air quality monitoring reports where the original data source was the U.S. Embassy station, and the air quality level posted was *bu jiankang* or "unhealthy", suggesting that censors may have viewed the juxtaposition of U.S. Embassy data on Weibo and the "unhealthy" air quality levels as sensitive. Although our findings regarding *physical harm* overall are null, this observation does suggest that even health-related posts can trigger higher censorship when linked to *political* content.

Finally, we consider our control variables for News and the real-time AQI. News does not seem to be affecting censorship much before June 5, with coefficients inconsistently signed across lags and insignificant or only weakly significant. Meanwhile, the real-time AQI at lag zero positively and significantly predicts increased censorship, while lag one predicts decreased censorship. This may suggest that censors rapidly restrict comments on high-pollution days, but then release control (potentially recognizing the impact of *visible censorship cost*). For the *Business as usual* phase, the overall takeaway is that government censors block *political* sentiment or allow more *scientific* commentary with a lag of one day. Next, we compare these results with those of the *Adaptive* phase in Table 5.

Instead of being significant at lag one, the majority of significant results in the *Adaptive* phase occur without a lag. For example, the signs and effect for Anti-Government are similar to Table 4, only this time at lag zero instead of one. The consistency of this measure across both time periods supports our theory that direct criticism of the government lowers its *credibility payoff* to not censoring, even with high *visible censorship cost*. For

Table 5: Censorship During the Adaptive Phase (Average Marginal Effects)

DV: Cens. Rate	Model I	Model II	Model III	Model IV
L.Cens. Rate	0.515***	0.462***	0.444***	0.417***
Political				
Dom. v. For.	-0.043	-0.031	-0.033	-0.065
L.Dom. v. For.	0.067***	0.045*	0.044*	0.000
Anti-Govt	0.105**	0.095**	0.080**	0.102***
L.Anti-Govt	0.042	0.019	-0.007	0.004
U.S. Embassy	0.438	0.555*	0.564*	0.567*
L.U.S. Embassy	-0.084	-0.118	-0.106	-0.126
Scientific				
AQI Monitoring	-0.183***	-0.212***	-0.192***	-0.205***
L.AQI Monitoring	-0.049	0.016	0.017	0.011
PM 2.5	-0.427**	-0.422***	-0.403***	-0.420***
L.PM 2.5	0.029	0.017	0.011	-0.024
Controls				
News		0.322***	0.318***	0.312***
L.News		-0.165**	-0.178**	-0.166**
Real-time AQI			0.036	0.031
L.Real-time AQI			-0.034	-0.041
Physical Harm				
Health				0.033
L.Health				0.043
Jiankang				0.056
L.Jiankang				0.019

$$^{*}\ p < 0.1\ ^{**}\ p < .05\ ^{***}\ p < .01 \qquad N = 200$$

the other *political* measures, U.S. Embassy is now positive, mostly significant, and much larger. This shows a clear distinction with the *Business as usual* phase, suggesting the government's attempt to shut down Embassy-related discussion after June 13. Domestic vis-a-vis Foreign is now positive and significant in Models I-III at lag one. Although not as strong as the result for Anti-Government sentiment, this may also be evidence of the state's determined effort to control *political* discussion after June 13.

Furthermore, the coefficients for AQI Monitoring and PM 2.5 are highly significant, negative and large at lag zero. This immediate and strong relationship between surges in PM 2.5 discussion and relatively lower censorship suggests a divergence in how the

government treated *scientific* versus *political* sentiments after June 13. With regard to AQI Monitoring, since many local governments opened air monitoring stations in the latter half of the year, lower censorship could highlight actual government action on air pollution concerns. Finally, in contrast to Table 5, all *physical harm* measures are insignificant. Although there might be evidence here for the neutrality of *physical harm*, there is also room for error in these measurements.

For the controls, News shows positive, significant and large coefficients across Models II-IV. While news content itself is unlikely to increase censorship since it is under close state supervision, dates with large amounts of news may provide fuel for criticism, which could prompt the censors to react to any relevant news story. Finally, real-time AQI is insignificant and small at both lags zero and one, suggesting that censors were not reacting to real-world pollution levels.

Conclusion

Overall, our study of Internet censorship in China supports the idea that autocrats sometimes relax control to signal government responsiveness. We provide evidence that censorship dropped during a "topic burst" in air pollution discussion on Weibo in June 2012, marking a *signal shift* in the government's response. We also show strong statistical evidence of variation in the state's approach to censorship with respect to different framings of air pollution. Directly disparaging *political* comments are the most likely to trigger censorship, while *scientific* comments consistently predict reduced censorship. Furthermore, a comparison of the period before (the *Business as usual* phase) and after (the *Adaptive* phase) the June events reveals stronger and more rapid effects for both sentiments after the crisis. Meanwhile, posts about *physical harm* do not strongly correlate with increased or decreased censorship in either period. Although this could be due to measurement error, it also suggests that the regime considers health-related commentary neutral or medium risk. This may explain the censors' lukewarm response to these remarks, but their decisive efforts with respect to *political* (high risk) and *scientific* (low

risk) comments.

Our case study illustrates the plausibility that our four factors (*responsiveness benefit*, *image harm*, *visible censorship cost*, and *collective action risk*) shape the incentives for the government's approach to information control. During a crisis situation, when the nation is paying attention to the state's actions and *visible censorship cost* is high, non-censorship introduces the possibility for *responsiveness benefit*, while harsh censorship ensures at least some amount of *image harm* or even *collective action risk* for appearing unresponsive. By not censoring at this time, the government might be signaling responsiveness, but in order to receive the full rewards of *responsiveness benefit* the government would have to follow through with its signaled policy change. Unless the government follows through with its implicit promise, when the next crisis hits, public anger will again erupt.

While we cannot prove that public anger causes policy change, we can consider the possibility by observing the government's actions after June 2012 with respect to air pollution. Since 2012, the Chinese government has stepped up transparency on environmental pollution through its data disclosure initiatives. By the beginning of 2013, the government had set up over 500 PM 2.5 monitoring stations in more than 70 cities around the country (D. Roberts, 2015). The following year, the government required 15,000 factories to publicly report real-time emissions data (Denyer, 2014). With each passing year, the government has released more data and announced sweeping initiatives to tackle the issue of air pollution, including declaring a "war on pollution" at the National People's Congress in 2014 (Reuters, 2014). It is possible that public pressure on Weibo played a role in accelerating these air pollution policies. For example, real estate mogul and outspoken Weibo blogger Pan Shiyi's calls for the government to be more transparent with PM 2.5 data were widely re-tweeted around June 13, contributing to the "topic burst" on that date. One high-profile Beijing-based environmentalist credited the Weibo discussion, and Pan Shiyi's role specifically, as major driving forces in the government's subsequent release of air pollution data with the PM 2.5 measure.[12]

The link between surges in online anger and expectations of government responsive-

ness and eventual policy change is further supported by recent research in Chinese politics. Some portion of Chinese citizens do expect the government to respond to citizen feedback channeled through these informal means. For those that do, citizens are less likely to comply with government directives that they see as misguided or inappropriate (Tsai 2015). Furthermore, citizens who observe harsh censorship perceive the government as less responsive to their demands and are more likely to take action through other means (Meng et al. 2017). Another study (Huang 2015) finds that the government cannot fully recover lost trust without a high-quality and strong rebuttal. Otherwise, even *rumors* of government inadequacies could erode political support. These studies suggest that 1) citizens in authoritarian regimes do expect the government to be responsive and 2) that the regime faces long term costs to legitimacy or compliance if they do not respond appropriately. This grounds the idea of *responsiveness benefit* in findings about how citizens in authoritarian regimes understand their own government's responsiveness and what motivates authoritarian adaptation to be responsive to citizen demands (Noesselt 2014). Even as President Xi Jinping centralizes power and cracks down on civil society, these dynamics linking citizen expectations of government responsiveness and the consequences of government inaction have not disappeared.

Finally, our argument about signaling responsiveness has implications beyond China for how authoritarians manage information during popular crises. For centralized and highly capable authoritarian regimes, our four variable framework models the incentives that government actors face in deciding whether to relax or restrict control. Instead of a "safety valve" (Hassid, 2012), or a "controlled burn" (Lorentzen, in press), we provide evidence that autocrats can relax censorship during a crisis to communicate responsiveness to citizens. This suggests a mechanism through which public grievances can be acknowledged, addressed, and incorporated into policy change in authoritarian regimes.

Notes

[1]See also Lorentzen (in press), which focuses on this vertical versus horizontal information transfer.

[2]Historical PM 2.5 data (beginning in 2008 for Beijing) is available on StateAir, the U.S. Department of State Air Quality Monitoring Program website: www.stateair.net.

[3]This estimate is a function of the observed censorship rate and Zhu et al.'s (2013) expectation that 90% of censorship happens within an hour of posting. See the authors' online appendix for more details.

[4]When air pollution is high, however, censorship would be easier for Internet users to detect. In 2012, pollution correlates with lower censorship only during the winter months, when both pollution and *visible censorship cost* are high. Since air pollution levels are relatively low in June actual air pollution does not factor into the dynamics that we describe during the *signal shift*.

[5]Average pairwise agreement for all hand-coded categories is well above 90% except for health (71.9%). Additional information on coding and inter-coder reliability is available in the authors' online appendix.

[6]*Shiguan* can refer either to an embassy (*dashiguan*) or a consulate (*lingshiguan*). Commenters commonly abbreviated "U.S. Embassy" (*meiguo dashiguan*) to "*shiguan*". We use *shiguan* as basic search term, since a keyword search for its two characters "*shi*" and "*guan*" would also catch "*meiguo dashiguan*."

[7]For more details on the ReadMe estimates, see the authors' online appendix.

[8]AQI was unusually high for three days, January 17 (327), January 18 (402), and January 19 (428). For reference, the average AQI for the year was 91 and AQI measures from 201-300 are considered "Very Unhealthy" and 301+ are "Hazardous".

[9]King, Pan, and Roberts (2013) define a "volume burst" as any event more than 3 standard deviations above the long-term mean. For January 19, the volume of posts was +3.9 sigma. For June 6, it was +6.2 sigma, and for June 13, it was +10.7 sigma. June 28 and 29 are also high volume dates, but are unrelated to the U.S. Embassy controversy, so they are excluded.

[10]We specify the lag order by using the Akaike Information Criterion (AIC), which led us to choose a lag order of four.

[11]The regressions in Tables 4 and 5 are run with lags two through four, but we are interested in only the most recent lags' effect on censorship.

[12]Interview by first author with an environmentalist at a domestic ENGO in Beijing, March 10, 2016.

References

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

Blaydes, L. (2011). *Elections and distributive politics in Mubarak's Egypt.* New York: Cambridge University Press.

Boix, C. & Svolik M. (2013). The foundations of limited authoritarian government: Institutions, commitment, and power-sharing in dictatorships. *The Journal of Politics, 75*(2), 300-316.

Bradsher, K. (2012, June 5). China asks other nations not to release its air data. *The New York Times*. Retrieved from https://goo.gl/MhlZIs

Brownlee, J. (2007). *Authoritarianism in an age of democratization.* New York: Cambridge University Press.

Chan, C. K. & Yao, X. (2008). Air pollution in mega cities in China – A review. *Atmospheric Environment, 42*(1), 1-42.

Denyer, S. (2014, February 2). In China's war on bad air, government decision to release data gives fresh hope. *The Washington Post*. Retrieved from goo.gl/nIMY3P

Dimitrov, M. K. (2014a). Tracking public opinion under authoritarianism: The case of the Soviet Union during the Brezhnev era. *Russian History, 41*, 329-353.

Dimitrov, M. K. (2014b). What the party wanted to know: Citizen complaints as a 'barometer of public opinion' in communist Bulgaria. *East European Politics and Societies and Cultures, 28*(2), 271-295.

Dimitrov, M. (2015). Internal government assessments of the quality of governance in China. *Studies in Comparative International Development, 51*(1), 50-72.

Egorov, G., Guriev, S., & Sonin, K. (2009). Why resource-poor dictators allow freer media: A theory and evidence from panel data. *The American Political Science Review, 103*(4), 645-668.

Esarey, A. (2013). Understanding Chinese regime censorship and preferences. Paper presented at the American Political Science Association Annual Meeting, September 1.

Esarey, A., & Xiao, Q. (2011). Digital communication and political change in China. *International Journal of Communications, 5*, 298–319.

Esarey, A., & Xiao, Q. (2008). Political expression in the Chinese blogosphere: Below the radar. *Asian Survey, 48*(5), 752–772.

Fu, K., Chan, C.H., & Chau, M. (2013). Assessing censorship on microblogs in China: Discriminatory keyword analysis and impact evaluation of the 'Real Name Registration' policy." *IEEE Internet Computing, 17*(3), 42-50.

Gandhi, J. (2008). *Political institutions under dictatorship.* New York: Cambridge University Press.

Gandhi, J., & Lust-Okar, E. (2009). Elections under authoritarianism. *Annual Review of Political Science, 12*, 403-422.

Gandhi, J., & Przeworski, A. (2006). Cooperation, cooptation, and rebellion under dictatorships. *Economics & Politics, 18*(1), 1-26.

Gehlbach, S., & Sonin, K. (2014). Government control of the media. *Journal of Public Economics, 118*, 163–71.

Hassid, J. (2012). Safety valve or pressure cooker? Blogs in Chinese political life. *Journal of Communication, 62*, 212-230.

Henochowicz, A. (2012). For better air, don't pin your hopes on embassies. *China Digital Times*. Retrieved from http://chinadigitaltimes.net/2012/06/for-better-air-dont-pin-your-hopes-foreign-embassies/

Hildebrandt, T., & Turner, J. L. (2009). Green activism? Reassessing the role of environmental NGOs in China. In J. Schwartz & S. Shieh (Eds.), *State and society responses to social welfare needs in China: Serving the people* (pp. 89-110). New York, NY: Routledge.

Ho, P. (2001). Greening without conflict? Environmentalism, NGOs and civil society in China. *Development and Change, 32*(5), 893-921.

Ho, P., & Edmonds, R. L. (2008). *China's embedded activism: Opportunities and constraints of a social movement*. London, UK: Routledge.

Hopkins, D. J., & King, G. (2010). A method of automated nonparametric content analysis for social science. *American Journal of Political Science, 54*(1), 229-47.

Huang, H. (2015). A war of (mis)information: The political effects of rumors and rumor rebuttals in an authoritarian country. *British Journal of Political Science, 47*, 283-311.

King, G., Pan, J., & Roberts, M.E. (2013). How censorship in China allows government criticism but silences collective expression." *American Political Science Review, 107*(2): 326-343.

King, G., Pan, J., & Roberts, M.E. (2014). Reverse-engineering censorship in China: Randomized experimentation and participant observation." *Science, 6199*(345), 1-10.

Kuran, T. (1995). *Private truths, public lies: The social consequences of preference falsification*. Cambridge, MA: Harvard University Press.

Little, A. (2016). Communication technology and protest. *Journal of Politics, 78*(1), 152-166.

Lohmann, S. (1994). The dynamics of information cascades: The Monday demonstrations in Leipzig, East Germany, 1989-91. *World Politics, 47*(1), 42-101.

Lorentzen, P. (2013). Regularizing rioting: Permitting public protest in an authoritarian regime. *Quarterly Journal of Political Science, 8*(2), 127-158.

Lorentzen, P. (2014). China's strategic censorship. *American Journal of Political Science, 58*(2), 402–414.

Lorentzen, P. (in press). Chapter 1: Introduction. In *China's Controlled Burn: Information management and state-society relations under authoritarianism.* Under contract at Cambridge University Press. Draft first chapter available online: peterlorentzen.com

MacKinnon, R. (2008). Flatter world and thicker walls? Blogs, censorship, and civic discourse in China. *Public Choice, 134*, 31-46.

MacKinnon, R. (2012). *Consent of the networked: The worldwide struggle for Internet freedom.* New York, NY: Basic Books.

Magaloni, B. (2006). *Voting for autocracy: Hegemonic party survival and its demise in Mexico.* New York: Cambridge University Press.

Magaloni, B. (2008). Credible power-sharing and the longevity of authoritarian rule. *Comparative Political Studies, 41*(5), 715-741.

Meng, T., Pan, J., Hobbs, W., & Roberts, M. E. (2017). Censorship of criticism reduces perceptions of government responsiveness. Unpublished paper.

Morozov, E. (2011). *The net delusion: The dark side of Internet freedom.* New York, NY: PublicAffairs.

Noesselt, Nele. 2014. Microblogs and the adaptation of the Chinese party-state's governance strategy. *Governance, 27*(3), 449-468.

Reuter, O.J., & Szakonyi, D. (2015). Online social media and political awareness in authoritarian regimes. *British Journal of Political Science, 45*(1), 29–51.

Reuters. (2014, March 4). China to 'declare war' on pollution, premier says. Retrieved from http://www.reuters.com/article/us-china-parliament-pollution-idUSBREA2405W20140305

Roberts, D. (2015, March 6). Opinion: How the U.S. Embassy tweeted to clear Beijing's air. *Wired.* Retrieved from https://www.wired.com/2015/03/opinion-us-embassy-beijing-tweeted-clear-air/

Roberts, M. E. (2014). Fear or friction? How censorship slows the spread of information in the digital age (Unpublished doctoral dissertation). Harvard University, Cambridge, MA.

Roberts, M. E. (2015). Experiencing censorship emboldens Internet users and decreases government support in China. Unpublished paper.

Second Author (2016). [unpublished paper; title omitted to preserve anonymity].

Second Author & Co-author (2016). [title and journal omitted to preserve anonymity].

Svolik, M. (2012). *The politics of authoritarian rule*. New York: Cambridge University Press.

Svolik, M. (2012) The foundations of limited authoritarian government: Institutions, commitment, and power-sharing in dictatorships. *The Journal of Politics, 75*(2), 300-316.

Tsai, L. (2015). Constructive noncompliance. *Comparative Politics, 47*(3), 253-279.

Wang, S. H. & Peng, M., (2015). Petition and repression in China's authoritarian regime: evidence from a natural experiment. *Journal of East Asian Studies, 15*(1), 27-67.

Whitten-Woodring, J., & James, P. (2012). Fourth estate or mouthpiece? A formal model of media, protest, and government repression. *Political Communication, 29*(2), 113-136.

Wintrobe, R. (1998). *The political economy of dictatorship*. Cambridge, UK: Cambridge University Press.

Yang, G. (2013). Contesting food safety in the Chinese media: Between hegemony and counter-hegemony. *The China Quarterly, 214*, 337-355.

Yang, G. (2009). *The power of the Internet in China: Citizen activism online*. New York, NY: Columbia University Press.

Yang, G., & Calhoun, C. (2007). Media, civil society, and the rise of a green public sphere in China. *China Information, 21*, 211-236.

Zheng, N. (2007). *Technological empowerment: The Internet, state, and society in China*. Stanford, CA: Stanford University Press.

Zhu, T., Phipps, D., Prigden, A., Crandall, J.R., & Wallach, D.S. (2013). The velocity of censorship: high-fidelity detection of microblog post deletions. eprint arXiv:1303.0597. Retrieved from Cornell University Library arXiv.org: http://arxiv.org/abs/1303.0597