Why autocrats sometimes relax online censorship of sensitive issues: A case study of microblog discussion of air pollution in China

Christopher Cairns[*]and Elizabeth Plantan[†]

October 9, 2016

Abstract

Air pollution is one of the most visible environmental problems in China, generating considerable online public discontent. However, despite potential social instability, Chinese leaders sometimes relax censorship over this discussion. Drawing on existing work in authoritarian politics, we theorize incentives for adjusting censorship and apply these to a dispute between the U.S. Embassy in Beijing and the Chinese government over the disclosure of air quality data. Using data from China's microblog, Sina *Weibo*, and relying on hand-coded and computer-assisted text analysis, regression models and a brief qualitative analysis, we find that Chinese leaders adapted to the crisis by either not raising, or lowering censorship in two categories – human health concerns and scientifically-grounded claims – while repressing direct criticism. In doing so, they signaled responsiveness to public concerns regarding pollution, but also re-delineated the tolerable limits of criticism. Our findings suggest sophisticated authoritarian adaptation to public demands through new technologies.

Keywords

China, social media, censorship, environment, air pollution, *Weibo*, responsive authoritarianism

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

[†]PhD Candidate, Cornell University. Email: enp27@cornell.edu

1 Introduction

In the wake of the Arab Spring, scholars have increasingly focused on the role of social media in authoritarian politics. Debate has centered on two aspects of new digital technologies: their role in facilitating anti-regime protest and their potential to be coopted by authoritarian governments as instruments of regime power. Scholars focusing on the first aspect (Howard & Hussain, 2011; Lynch, 2011; Shirky, 2011; Tufecki & Wilson, 2012; Reuter & Szakonyi, 2015; Farrell, 2012; Castells, 2012; Bennett & Segerberg, 2013) have highlighted social media's ability to facilitate anti-regime coordination. Work focused on the second aspect, in contrast, has stressed savvy autocrats' ability to harness new media through multiple means, such as online information-gathering about citizen discontent and using social media for propaganda purposes (Aday, Farrell, Lynch, Sides, Kelly, & Zuckerman, 2010; Aday, Farrell, Lynch, Sides, & Freelon, 2012; Morozov, 2011; MacKinnon, 2012; Gehlbach & Sonin, 2014; Dimitrov, 2014a; 2014b; Lorentzen, 2014).

These two aspects are not mutually exclusive, but the circumstances under which protesters or the state gain the advantage online remain underexplored. What is certain, however, is that autocrats are paying increased attention to controlling such digital forums. While the overall trend has been toward increased control, a growing body of formal theory (Egorov, Guriev, & Sonin, 2009; Whitten-Woodring & James, 2012; Gehlbach & Sonin, 2014; Lorentzen, 2014; Little, 2016; Reuter & Szakonyi, 2015) has shown that full media censorship is not optimal for autocratic survival. Rather, under some circumstances, *relaxing* control may yield benefits to the state. Such a temporary opening of the digital floodgates has occurred even in China, where social media is tightly controlled. In 2012, nationalist protesters were briefly allowed to vent their anger on *Weibo*, the country's leading microblog, toward their own government for not 'standing tough' against Japan with regard to disputed islands in the East China Sea (First Author & Co-author, 2016). And in 2015, social media discussion of the environmental documentary *Under the Dome* was allowed to proliferate for almost a week before the film was pulled offline and commentary censored.

This paper offers a unique answer to why leaders would temporarily relax control over social media. Existing theories (such as Lorentzen, 2014) have proposed that slightly looser control enables rulers to gather information about the sources of discontent. While this explanation has merit, we argue that it cannot explain cases of conspicuous non-censorship that occur during periods of heightened public awareness. Rather, we suggest that temporary openness in these instances serves to persuade the social media-using public that the regime acknowledges their concerns, a deliberate move by the state that we term its *responsiveness benefit*. We illustrate such a possibility with a case study of online discussion about China's air pollution problem during 2012, a year in which leaders clashed with the U.S. Embassy in Beijing over the latter's online publication of its own air monitoring data that showed the problem to be more severe than Chinese data had indicated.

As the controversy ebbed and flowed throughout the year, different online sentiments emerged, with some commentators focused on pollution's threat to human health, others adopting a more scientific approach, and still others lambasting the government as ultimately responsible. Through a combination of hand-coded and computer-assisted content analysis, as well as statistical modeling of sentiment trends' temporal relation with *Weibo* post deletions, we show how censorship varied over time depending on which of these three sentiments predominated on any given day. Additionally, we find that censorship across all three sentiments was relatively high early in the year, but diverged after Chinese authorities' dispute with the U.S. Embassy peaked in June, with more hostile sentiments (toward the government) being censored more tightly thereafter and more neutral sentiments more loosely; we find such a trend consistent with leaders' greater need to show responsiveness on the air pollution issue, but also greater fear of online hostility.

Such a complex pattern suggests that, at least in this instance, leaders pursued a highly nuanced online censorship response. Disaggregating online censorship in this manner allows us a dynamic and fine-grained look at the state's ability to micro-tune information control according to various incentives. Although our findings suggest that leaders relax censorship for certain sentiments and during times in which the state's *re-*

sponsiveness benefit is likely to be greatest, elites also take into account some of the risks entailed by information openness, such as the threat of collective action (King, Pan, & Roberts, 2013; 2014) and the potential for broader damage to the Party's (and their personal) reputations (Esarey, 2013). A fourth factor is the potential backlash leaders face if the social media public becomes aware they are being censored, since citizens think the state is trying to hide bad news (Roberts, 2015).

In considering the latter three factors, we acknowledge their utility in explaining why online censorship in China does (or does not) occur in specific instances. Each is grounded in extant literature and is not mutually exclusive with our account. The paper's main contribution is therefore to add *responsiveness benefit* as a critical missing piece that enriches our knowledge of online censorship both in China and in other authoritarian states. A second contribution is that our findings are inconsistent with theories proposing that Internet information openness is always about gathering information. While not negating the fundamental information challenge all authoritarian rulers face and their need to develop media (or other) institutions to deal with the problem, this paper suggests that at least for high-profile incidents on social media, when autocrats choose not to censor they are interested not in incoming information flows from society, but in outward communication: implicitly using bloggers' own voices to show responsiveness to growing popular demands.

2 Relevant literature

Much work on authoritarian information control has been framed around 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 are afraid of looking disloyal. As Kuran (1995) describes, citizens in authoritarian regimes 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. Authoritarian leaders, then, often lack information about their true level of support, public grievances, or the impact of policies. To address this deficit, leaders have used multiple channels, including public opinion polls and formal citizen complaints (Dimitrov, 2014a; 2014b). Yet such methods can backfire if the information-gathering means used in fact increase citizens' "common knowledge" (Kuran, 1991; 1995) of each others' discontent and have the potential to enable regime-toppling "information cascades" (Lohmann, 1994; 2000).

Many scholars have used formal models to consider this tradeoff between the need for information and the risk of spreading popular awareness of shared discontent (Egorov, Guriev, & Sonin, 2009; Whitten-Woodring & James, 2012; Gehlbach & Sonin, 2014). However, these studies often only give the government a dichotomous option between repression and liberalization and focus on traditional rather than new media sources. Other scholars (Little, 2016; Reuter & Szakonyi, 2015) have considered online content and social media, but their focus is on the Internet's implications for collective action, rather than more complex (and less existentially threatening) online interactions between state and society. These studies also focus more on the conditions under which information would be censored, rather than considering the choice not to censor. But as Dimitrov and Sassoon (2014) argue, "totalistic repression exists in information-poor environments," and that "as the quality of information increase[s], repression becomes more selective and targeted" (p. 3). Within a more targeted policy of censorship, there will be times in which control is more relaxed. Shadmehr and Bernhardt (2015) find that when deciding whether or not to censor media reports critical of the regime, authoritarian rulers gain by censoring slightly less. Why would a capable regime choose not to fully repress online information?

Within the context of China, many scholars have considered possible interactions between the Internet, public discontent, and state censorship. Studies have focused on why the state would censor: fear of collective action (King, Pan, & Roberts, 2013; 2014); desire to preserve top leaders' reputation (Esarey, 2013); and the ease with which the state is able to make censorship invisible to netizens, thus lowering its cost (Roberts, 2014). Many other studies (Yang, 2009; Zheng, 2007; MacKinnon, 2008; Morozov, 2011) have illustrated the high degree of control Chinese leaders exercise over the Internet, but have not adequately addressed areas of more nuanced control or considered a more dynamic conception of varying censorship in response to strategic circumstances. Non- or relaxed censorship, as observed by one of the authors (First Author & Co-author, 2016), has received less attention and deserves further development. Until now, scholars have mostly considered relaxed censorship as due to autocrats' information-gathering need. Such work has not considered the possibility, however, that more 'public' information-gathering channels like social media are two-way streets between state and society. Not only can the state use social media to gather information, but we contend that the online public will draw an inference about the state's commitment (or lack thereof) to meet citizen demands based on how tightly the state censors that particular topic.

This idea, that temporary social media openness could send a *persuasive signal* to citizens about the state's plans, differentiates our explanation from others that also address the act of censorship as medium for state-society interaction. Hassid (2012) has posited that allowing online grievances acts as a "safety valve" for citizens to vent their feelings. Lorentzen (in press) argues that the situation is more of a controlled burn in which foresters permit fires to break out or even light them in order to burn off this flammable material in a contained area (p. 9). In both metaphors, the state permits some airing of discontent at some risk to itself (since safety valves can explode and controlled burns can turn into wildfires), in order to pre-empt worse problems later on, but neither considers that online citizens might then draw some inference about state capabilities and intentions from the more open environment. In contrast, we view non- and lowered censorship more strategically as a signal of responsiveness from leaders to online citizens. In a country (China) where stringent social media censorship is the norm, periodic non-censorship on politically sensitive topics, especially when voices criticize the state for not addressing the problem, constitutes a 'deafening silence' that speaks volumes about leaders' acknowlegement of public discontent.

3 Analyzing Internet censorship: a four-variable framework

Before beginning analysis of the factors that influence authoritarian states' online censorship decisions, we need to briefly address their *capacity* to actually implement a desired censorship level. In this respect, robust censorship can be viewed as a policy area like many others – enjoying wide consensus among top leaders as to what should be done, but facing challenges in inducing bureaucrats to achieve the objective. Current theories of policy implementation in China often adopt one of two frameworks for studying the bureaucracy. The first assumes that the state is centralized, with rational decision-makers who censor with "military-like precision" (King, Pan, & Roberts, 2013, p. 1). This literature has also empirically established that censorship is extremely rapid and effective in response to breaking incidents (Zhu, Phipps, Prigden, Crandall, & Wallach, 2013; King, Pan, & Roberts, 2013; 2014; Bamman, O'Connor, & Smith 2012). The second framework falls under the category of "fragmented authoritarianism" (Lieberthal & Oksenberg, 1988; Lieberthal & Lampton, 1992) and unpacks the state into decentralized units. These studies focus on multiple censorship-relevant agencies with overlapping jurisdictions, or look for examples of Internet companies that resist or delay orders to delete content. Although we acknowledge the importance of both approaches, we assume a centralized approach to censorship. Recent work by one of the authors (First Author, 2016) has shown that Internet censorship is reasonably centralized and responsive to central directives.

On this basis, we argue that the Chinese state can be treated as unitary, top-down and disciplined for purposes of censoring social media and that four rational cost-benefit factors strongly shape (although they do not completely determine) online censorship decisions: *responsiveness benefit*, *image harm*, *visible censorship cost*, and *collective action risk*. While the latter three are rooted in existing work, as mentioned earlier, the concept of *responsiveness benefit* is new. It is the benefit the government receives in netizen eyes by acknowledging that some problem exists. By allowing bloggers to talk about the issue, the government implicitly signals acknowledgment that a problem exists *and* has prompted legitimate citizen concern. Conversely, if netizens expressed concern about a problem but observed swift censorship, they would infer that leaders either did not view the issue as a problem (unlikely in the case of air pollution), or that they recognize the problem, but consider the public airing of grievances about it politically unacceptable. This logic differs from Dimitrov, Lorentzen, and others in that *responsiveness benefit* is an *interactive* form of "responsive authoritarianism" in which the state acknowledges popular grievances and communicates with society; in this respect, the concept is more akin to various scholars' work on the media's role in "public opinion supervision" (Esarey & Xiao, 2011; Stockmann, 2013), i.e. bottom-up pressure that results in the state seeking to demonstrate responsiveness.[1]

The logic of *responsiveness benefit* presumes that netizens both observe when censorship occurs and interpret non-censorship as a signal intended by leaders not only to appear responsive to public demands, but *strong* and *capable* of immediately dealing with the problem. By allowing discussion, China's elites know that citizens will generate collective perceptions of leaders' responsibility to fix the problem, including some amount of criticism. As former Paramount Leader Deng Xiaoping stated, "if you open the window for some fresh air, you have to expect some flies to blow in." Of course by relaxing censorship on the issue, some fraction of online participants may sharply blame elites or the Party for the problem. But leaders may hope that the majority of online observers will perceive even limited openness as a signal that real reform is just around the corner. Indeed, in the history of Chinese Communist rule, previous episodes of briefly tolerated free speech on select topics have served as a prelude to reform, the most famous example being the "Democracy Wall" of 1978-79 which presaged major 1980s reforms.

For savvy regimes that hope to survive, such spaces are not tolerated indefinitely because of their potential to spiral into broad anti-regime activities. Even if actual collective action is unlikely (as is often the case for online speech), however, a less acute threat exists if leaders open up. The framework's second factor, *image harm*, is defined

[1] To be clear, relative online openness does not constitute true responsiveness to 'mass' opinion; social-media using citizens, particularly on microblogs like Sina *Weibo*, tend to be urban, highly educated and wealthier than the average citizen.

as leaders' estimation of the probability that a majority of netizens 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 citizens are unwilling or unable to physically mobilize in the short term, inferring state weakness or ineptitude (and importantly, generating collective knowledge of this fact through online discussion) increases their potential for anti-regime activity or resistance down the road. In a sense, *responsiveness benefit* and *image harm* can be viewed as two sides of the same coin, where one entails citizens' shared perceptions of regime strength, and the other, weakness. We thus define a composite factor, *credibility payoff*, according to Equation 1:[2]

$$credibility\ payoff = responsiveness\ benefit - image\ harm \tag{1}$$

The third factor, *visible censorship cost*, addresses citizen awareness of being censored. Research (Roberts, 2014; 2015) has shown that when censorship is sufficiently invisible – that is, netizens may suspect they are being censored but are not sure and are unable to find any relevant information through searching or browsing – they give up seeking the information. On the other hand, if individuals *know* they are being censored, Roberts finds that they are actually more motivated to post sensitive information. Therefore, highly visible censorship during high-profile political events can backfire, increasing pressure on the government. Whether this occurs will depend on the episode specifics, the public's pre-existing awareness of the problem, and the availability of alternative information sources (such as Chinese or foreign media reporting). Thus, the government's *expectation* that a particular censorship act will be too visible, and belief that leaders will pay a price in terms of public reputation, can motivate their decision to refrain from cracking down.

The fourth and final factor, *collective action risk*, has received substantial empirical support in the literature on censorship in China (King, Pan, & Roberts, 2013; 2014).

[2] We think it is reasonable to take the difference of *responsiveness benefit* and *image harm* to form a single measure because they both fundamentally concern the inference leaders think the *Weibo* public will draw *if an episode is not censored.* That is, will more people interpret online openness as positive government acknowledgment of a problem, or as weakness?

Social media posts during so-called “topic bursts” (surges in online discussion surrounding a specific incident or topic) that relate to real-world collective actions such as street protests are more likely to be censored. We accept this finding as an important factor in determining much Internet censorship in China. However, authorities’ fear of collective action is unlikely to explain the censorship pattern regarding the 2012 U.S. Embassy dispute for the simple reason that no street protests or other forms of real-world coordination occurred.[3] While Chinese citizens have taken to the streets to protest other environmental threats such as the construction of chemical factories (Lang & Xu, 2013; Chen, 2009), to our knowledge no such mobilizations have occurred in response to spikes in air pollution levels. We therefore do not think that collective action risk accounts for the censorship we observe.

3.1 Applying the framework: “static” and “adaptive” phases in China’s social media management

In applying the framework to online incidents, we begin with the observation that in China, contested issues of relevance to both the state and the social media public tend to evolve along a repetitive path from when they are first mentioned. While in China the majority of online topics are not considered ‘political’ and not all political topics are ‘blacklisted’ *ex ante*, for those that are, the in-house censors at Internet companies have lists of banned keywords and are supposed to immediately delete any topic containing those words; in the life cycle of a social media topic, we term this period the state’s “Static Phase”.[4] Topics that truly engender sustained online public interest can sometimes survive by netizens altering the words and phrases they use, making censorship more difficult. While censors often do their best to repress a banned topic in its early

[3] We confirmed this with a LexisNexis search for any foreign (English-language) media reporting of air pollution-related protests during key dates in 2012, which yielded zero results. We relied on foreign media since these are not subject to the same in-house censorship bias as Chinese outlets. While foreign media’s coverage of Chinese protests is often spotty, any protest in major cities large enough to be of real concern to authorities would have been covered by international media.

[4] We chose the word “static” to connote a period in which the state’s censorship response is expected to be “business as usual”, i.e. rigid, conservative and in accordance with previously established procedures (where these exist). “Static” also contrasts to what we view as the state’s more dynamic and adaptive response later on.

stages, 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 on some issues leaders reach a turning point – an "Adaptive Phase" – where they see the benefit of opening up selective space for tolerable criticism, while more aggressively filtering destabilizing or de-legitimizing comments.

We argue that our four theoretical factors are a decent approximation of the shifting costs and benefits authoritarian leaders perceive for allowing or censoring certain types of online comments. However, while the factors themselves are broadly applicable both within China and potentially elsewhere, we maintain that they are best operationalized in a *context-specific* manner that takes into account the unique characteristics of different regimes, issues, and online spaces. In this paper we utilize a specific case (a dispute between the U.S. Embassy and the Chinese government over air pollution data) on one platform (Sina *Weibo*) during a specific time frame (the year 2012) to study these dynamics, and next briefly justify our choices.

3.2 Case selection and description

As an illustration of the framework, we have chosen to examine Internet censorship policies in China around a specific issue (air pollution) on a particular social media platform (*Weibo*). The discussion of air pollution on Weibo during this time period is a "most likely" case for illustrating strategic and adaptive censorship policies in China, an authoritarian regime with a sophisticated infrastructure of information control. We have chosen China as a case because of its documented ability to control information within its borders. This capacity for control has even inspired admiration among other authoritarian regimes. As Lorentzen (in press) notes, the Chinese Communist Party has become a "role model and standard-bearer" for other authoritarian leaders (p. 16). Second, we focus on environmental issues because their seriousness and scale have aroused public concern, yet discussion in this area remains relatively open (Ho, 2001; Ho & Edmonds, 2008; Yang & Calhoun, 2007; Hildebrant & Turner, 2009). Since environmental activism enjoys a privileged space in Chinese domestic politics, we would expect censorship of this issue to fluctuate between full repression and full liberalization.

In addition to choosing a relatively open issue area, we also selected a case that has arguably attracted the most attention from among the educated, Internet-using Chinese public: air pollution. Despite a willingness to address the problem, the Chinese government has struggled to control the narrative over air pollution management, particularly online. This struggle was exemplified by events in 2012, when the release of air quality monitoring data became a political flashpoint between the U.S. Embassy in Beijing and the Chinese government. This combination of distrust of government air quality measurements and widespread acceptance of the U.S. Embassy Beijing's data fueled a heated online discussion about air pollution, government responsibility, and foreign influence throughout the year. Finally, our study focuses on examining online discussion of this controversy on *Weibo*, China's version of Twitter. *Weibo*, which was at its peak as a lively public forum around 2012, is ideal to test the concept of *responsiveness benefit* since the temporary relaxing of censorship would be most obvious (and thus, most likely to strongly signal state reform intentions) in such a viral, public digital space. "Public squares" like *Weibo* are very effective at enabling "common knowledge" (Kuran, 1991; 1995) about the existence of some problem and who is responsible – precisely what we argue state leaders would want the public to infer in this case.

Before discussing data and methods, we first briefly discuss our case study and justify why we have chosen events in June 2012 as the demarcation between the "Static" and "Adaptive" phases in government censorship policy toward air pollution on Weibo.

3.2.1 Identifying the break point in the state's adaptive shift: June 6 and 13

Although daily Air Quality Index (AQI) data has been available in many Chinese cities since the early 2000s, a more recent controversy arose over the U.S. Embassy in Beijing 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). [5] On World Environment Day (June 5), after years of private complaints about the U.S. Embassy's release of its monitoring data, Ministry of Environmental Protection Vice-Minister Wu Xiaoqing finally went public, accusing 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. Netizen reactions were overwhelmingly negative and mocking of the government. Many commenters showed a general awareness that had it not been for the Embassy publicizing its data, no national online discussion of air pollution would have taken place.

Weibo commentary about Wu's remarks continued to simmer for several days after June 5 and 6 despite persistent government attempts at censorship. Just as it began to peter out, 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 the ones held accountable for improving the situation. The next day (June 13), netizen responses to Cui were even more mocking than the previous episode, with bloggers slamming him for attempting to divert blame away from what many viewed as a government cover-up.[6]

Because the dates of June 6 and 13 represent two very high-profile public comments by top Chinese officials – and also book-end the period of most lively and provocative *Weibo* activity in response to these comments – they are a natural turning point at which to look for the state's adaptive shift. We therefore argue that June 13 represented the peak of public mobilization, after which the government reasserted control over hostile speech, while accommodating perceived non-threatening framings of the issue. The following empirical analysis provides evidence for the shift that we have qualitatively described here.

[5] 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.

[6] We found in our data that the government *reduced* the amount of topic-relevant censorship on that day to an unusually low 30%, allowing a flood not only of less controversial speech about pollution and PM 2.5 to be widely viewed online, but even much more 'political' speech. While this striking anomaly is not our main focus (we are more interested in analyzing fluctuating censorship over the course of the year), it is consistent with other episodes such as nationalist protest (First Author & Co-author, 2016) where leaders have been observed to lower censorship at puzzling moments.

4 Data and method

To illustrate our arguments, we relied on a trailblazing dataset collected by researchers at the University of Hong Kong ("WeiboScope") that consists of over 38,000 *Weibo* celebrity users (Fu, Chan, & Chau, 2013), which the researchers defined as all users with verified identities as public figures and more than 10,000 followers as of January 2012. To our knowledge it is the most comprehensive dataset of *Weibo* posts currently available and the methodology used to collect it is described in detail in Fu, Chan, and Chau (2013). While such a sample is biased toward educated elites, as opinion leaders these individuals are in fact our population of interest since what they say has great potential to directly affect officials' perceptions of online opinion's danger or volatility. Each row in the dataset consisted of one social media post plus associated meta-data. We relied only on the post text, and counted 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. The dataset includes a timestamp for when a post was last publicly available and then is marked as "censored." While this method was not perfect, it is the best available method to get some measure of the speed and volume of censorship. However, some fraction of posts could be deleted prior to the WeiboScope program taking its daily record, which means that the actual rate of censorship may be much higher than the WeiboScope dataset suggests. To address the potential under-reporting of censorship and based on prior work (First Author & Co-author, 2016), we use a mathematical correction to estimate the "true" censorship rate from existing information, subject to some assumptions. This estimation is discussed in detail in Appendix B.

4.1 Operationalizing the framework: Predicting censorship by sentiment category

On the independent variable side, we operationalize our framework by deriving sentiment categories from a reading of the *Weibo* data, and studying the fluctuations of these categories in relation to changes in daily censorship as observable implications of changes in *credibility payoff* and *visible censorship cost* over time. Our method for coding the posts and their resulting sentiment categories is detailed in the next section. Here, we use those coded categories and match them to our predictions of how each particular category, in Chinese leaders' minds, would likely be associated with negative, neutral or positive *credibility payoff* and *visible censorship cost* at different points during 2012. In terms of its impact on censorship, we assume that *credibility payoff* dominates *visible censorship cost* and that the two vary according to different patterns, with the latter varying over time but equal across sentiment categories, and the former varying across both dimensions.

We score each independent variable – our three sentiment categories of political, physical harm, and scientific commentary – on a scale from -2 to +2, with -2 as "Very Negative", 0 as "Neutral" and +2 as "Very Positive". We weight *credibility payoff* as twice as important (2x) as *visible censorship cost*, and then sum the two scores to yield predicted censorship. The mathematical terms in the top lines of Tables 1 and 2 show each variable's signed relationship to censorship: both are *inversely* related to censorship, meaning that a higher score for each is associated with *reduced* censorship. We scale censorship from -6 to +6 (the equation's theoretical minimum and maximum values), with -6 representing a theoretical ideal of "Very Low" censorship, +6 representing "Very High" censorship, and 0 representing "Partial" censorship.[7]

[7] These scores obviously do not correspond to actual percentages of deleted/censored posts, since the long-term averages of these for all politically sensitive censored incidents on Chinese social media are not known. For reference, King, Pan and Roberts (2013) find an average for topics related to collective action of about 57%. Here, we do not intend to predict actual censorship levels but rather develop a categorical scheme to capture our variables' relationship with censorship's *relative* magnitude.

Table 1: Predicted Censorship by Sentiment Category ("Static Phase": January 2 – June 5)

Category	Cred. Payoff ($-2x$)	Vis. Cens. Cost ($-x$)	Pred. Censorship
Political	Negative (-1)	Neutral (0)	+2
Physical Harm	Neutral (0)	Neutral (0)	0
Scientific	Positive (+1)	Neutral (0)	-2

Table 2: Predicted Censorship by Sentiment Category ("Adaptive Phase": June 14 - December 30)

Category	Cred. Payoff ($-2x$)	Vis. Cens. Cost ($-x$)	Pred. Censorship
Political	Very Negative (-2)	Positive (+1)	+3
Physical Harm	Positive (+1)	Positive (+1)	-3
Scientific	Very Positive (+2)	Positive (+1)	-5

First, we coded *visible censorship cost* as "neutral" for the Static Phase because we had no reason to believe that the U.S. Embassy dispute would be atypically easy or difficult for authorities to cover up. Although real-world crises like natural disasters can increase *visible censorship cost*, officials' own statements and state media reporting typically play a larger role – and such statements were absent from January-June. Such a situation differed from the Adaptive Phase (after June 13), where we coded *visible censorship cost* as "Positive" (+1); on the one hand, the crisis dates of June 6 and 13 likely had a significant impact on raising public awareness of the issue, but on the other, such awareness would tend to diminish over time as *Weibo* users' attention shifted elsewhere.

Turning to *credibility payoff* in the Static Phase, we expect it to be "negative" for *Political*, "neutral" for *Physical Harm* and "positive" for *Scientific.* We expect *Political* to be negative because prior to the June dispute, officials likely saw no benefit (and some harm) in allowing any public comparison of China's own air monitoring data statistics to the U.S. Embassy data. The *Physical Harm* category, on the other hand, is somewhat less sensitive since citizen fears of pollution's health impacts, while possibly generating some pressure, are not as directly embarrassing for leadership as more political speech. Finally, the *Scientific* category is the least sensitive. This is because of the government's long-standing tolerance of public discussion backed by scientific data and their commitment as of January 2012 to establish new air monitoring stations nationwide. For the Adaptive Phase, we coded *credibility payoff* as "very negative" for *Political*, "positive" for *Physical Harm* and "very positive" for *Scientific*. Here, the three sentiment categories diverge as to positivity/negativity, since we argue that the central leadership in this phase decided to fully legitimize scientifically-rooted commentary, show some tolerance of worries about pollution's physical harm, and firmly crack down on politically sensitive speech.

These tables then provide the foundation for statistically-based inferences that link censorship to our theoretical framework. The tables predict *relative levels* of high or low censorship resulting from our independent variable scores. However, the data shows various *trends* (increases and decreases) in our sentiment categories over time. We can link predictions of censorship levels to these trends by treating the latter as observable implications of the former. Specifically, for individual measures (e.g. keywords or human-coded sentiments) that proxy for sentiment categories coded as "positive" for *credibility payoff*, particularly during the Adaptive Phase with "positive" *visible censorship cost*, increases in the proportion of all posts belonging to that category on a given day should lead to short-term *decreases* in the overall censorship rate. Conversely, increases in measured sentiment proportions for categories in which *credibility payoff* is negative (and especially during the "Static Phase") should lead to short-term *increases* in daily censorship.[8] These dynamic relationships should then be apparent in regression models linking daily censorship to our measures.

[8] An even more robust approach would be to measure changes in censorship *within* individual sentiment categories over time rather than overall daily censorship. However, obtaining reliable estimates of within-category censorship rates would require sub-sampling and coding many times more posts (many thousands instead of hundreds) as available time and resources permitted. Additionally, some dates simply do not have enough posts to obtain sufficient sample sizes for less frequent measures and categories.

4.2 Coding methods and measures

To filter out only pollution-relevant data, our sample consisted only of 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 posts for all of 2012. We went through several stages of pre-coding exercises to determine the key categories before moving on the full coded sample. Appendix A details our procedure. After several rounds of pre-coding exercises, we settled on our key measures. As mentioned earlier, these fit into three larger sentiment categories: 1) political criticism; 2) concerns about physical harm; and 3) scientific information. For the *Political* category, we included three measures. First, we wanted to capture the sentiment of Chinese comparing the air quality situation in their own country to other countries or to the international community. We termed this measure "Domestic vis-à-vis Foreign." Recent work by one of the authors (First Author & Co-author, 2016) has highlighted the prevalence of nationalist discourse on *Weibo* and the pervasiveness of Chinese citizens' view of themselves vis-à-vis other countries. For top leaders, this discourse is among the most difficult to manage of all political themes, since it questions the state's own legitimating narrative. While codings of domestic vis-à-vis foreign encompassed both pro- and anti-state commentary, we found that a large majority of such comments could be read as reflecting 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 this category simply "Anti-Government". Our third and final *Political* measure was the keyword "U.S. Embassy" (in Chinese) itself, which we found to proxy well for politically critical speech on the issue of air pollution in 2012.

For the *Physical Harm* category, we included several measures of whether air pollution-related comments framed the issue as a threat to human health. Since reliable coding decisions for this measure proved uniquely difficult, we also added an additional keyword measure "*Jiankang*", which is simply the Chinese word for health. 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. To capture a different but related scientifically-grounded speech trend, we also counted daily occurrences of the keyword “PM 2.5”. Although the term appeared in a variety of contexts, some of which overlapped with *Political* and *Physical Harm*, we chose it to represent *Scientific* 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 measured the presence of “News” in *Weibo* by counting all posts containing a left bracket (“[”) which nearly always signifies the beginning of a news story link. Our specific concern was that spikes in pollution-relevant news stories might both increase the prevalence of certain sentiment categories, and directly cause an increase in censorship as censors took the news media’s activity as a sign of an overall more volatile situation. This would confound estimation of the independent censorship effect of the category fluctuations themselves. An additional control consisted of actual air quality data taken straight from the Beijing U.S. Embassy’s rooftop monitoring station in 2012 (“AQI Index”); we included this measure to condition all of our results on real-world pollution fluctuations.

This exercise had two goals. First, we wished to estimate the proportions of posts in each category for June 6 and 13. After hand-coding a sample of 500 posts that spanned the whole year, we sub-sampled 150 posts from each of these two dates. While June 6 and 13 themselves are less the focus of our empirical testing than the dates before and after them, they do enrich our understanding of what led Chinese leaders to shift censorship strategy. Second, we aimed to generate year-long time series to chart the changes in our sentiment category proportions. However, since drawing and coding a post sample from each day of the year was infeasible, we used a computer assisted text analysis (CATA) algorithm called *ReadMe* (Hopkins & King, 2010) to estimate the proportions for the entire year. See Appendix C for more details.

5 Results

Before moving to regression modeling, we first present summary statistics and graphs of our estimated keyword and *ReadMe*-based proportions. Table 3 reports estimated mean proportions of all sentiment measures divided up into our four time periods.[9] For reference purposes, the average Air Quality Index (AQI) from the Beijing U.S. Embassy monitoring station is included.[10]

Table 3: Estimated Sentiment Measures and AQI Average

Measure	Jan 2 - Jun 5	Jun 6	Jun 13	Jun 14 - Dec 30
Domestic-vis-a-vis-Foreign	.22	.18	.83	.16
Anti-Government	.34	.23	.84	.21
U.S. Embassy (keyword)	.04	.25	.69	.02
Health	.28	.24	.66	.22
"Jiankang" (keyword)	.08	.03	.02	.07
AQI Monitoring	.29	.11	.25	.42
PM2.5 (keyword)	.38	.32	.12	.25
News ("[" measure)	.33	.28	.09	.37
U.S. Embassy AQI Index	97	143	70	87
Censorship Rate	.49	.71	.30	.64
Daily Average Posts	181	1460	2363	164[a]

[a] Numbers for Jun 14 - Dec 30 exclude June 28 and 29, which concerned an incident unrelated to the Embassy dispute that contained pollution-relevant keywords.

[9] The keyword measures are simply the count of each keyword over total posts for a given date or time period.

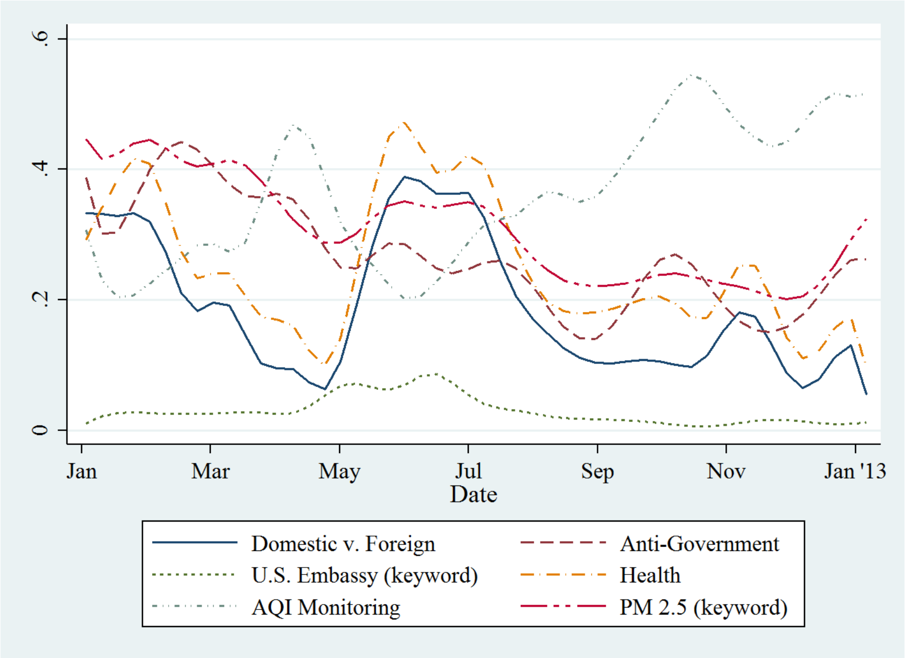
[10] The AQI is a composite measure of multiple pollutants, but is heavily influenced by ambient levels of PM 2.5. A 0-50 reading is considered "Good"; 51-100 "Moderate"; 101-150 "Unhealthy for Sensitive Groups"; 151-200 "Unhealthy"; 201-300 "Very Unhealthy" and 301+ "Hazardous".

From January 2 to June 5, the "Static Phase," a few proportions stand out: the PM 2.5 keyword was widespread, as was health-related commentary. In addition, the proportion of news stories was substantial (.33). When reading through the posts, we found that much of this news concerned local government initiatives to bring new air quality monitoring stations online. Yet however proactive these state-directed efforts might have been they did not seem to stem various *Political* criticisms (Domestic vis-a-vis Foreign and Anti-Government speech), which were substantial when compared with later in the year. Finally, the censorship rate, though not low in absolute terms, was lower (.49) than the year-long average (.57).

We next consider June 14 to December 30, which we argue represents the "Adaptive Phase" in censorship policy. Our *Political* measures showed marked declines compared with earlier in the year, particularly Anti-Government (.34 to .21). At the same time, News, and the *Scientific* category – notably AQI Monitoring – increasingly dominated the topic blend. In contrast, our *Physical Harm* variables were lower than previously but not as low as *Political.* Finally, the censorship rate showed a substantial increase (.64). Overall, these proportions suggest that what leaders perceived as less threatening sentiment categories became increasingly prevalent after June 13, while the more threatening *Political* category was increasingly restricted.

Third, we examine year-long graphs of our hand-coded and keyword proportion estimates in Figure 1.

Figure 1: Sentiment Measure Time Series

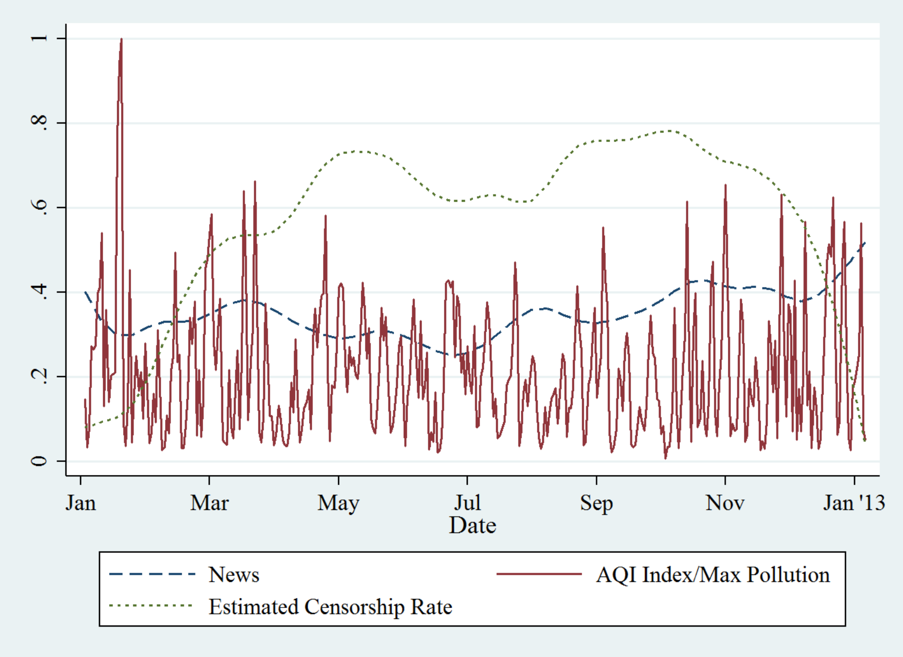


Looking at the graph, we notice a surprising finding: a strong correlation between Domestic vis-à-vis Foreign and Health, which we had expected to diverge since they were supposed to represent two different sentiment categories. The reason for such a correlation was not immediately evident, but became clearer upon a qualitative reading of the data. While during coding we treated Health holistically to include all manner of netizen concerns about pollution's harmful effects, in practice these concerns tended to increase alongside domestic-foreign comparisons, specificially references to World Health Organization air quality standards, an observation that we explore further in our regression results.

The other measures exhibit more independent variation. Anti-Government posts begin the year strong before gradually declining, except during the turning point in June. While not showing a sharp break between the pre- and post-June periods, this trend could signify a gradual shift toward tighter censorship of anti-government speech after June 13. U.S. Embassy posts, in contrast, mostly occur in the months prior to and during June 6-13, dropping to near zero thereafter. Mentions of PM 2.5 prevail early in the year, resurge during June, then gradually decline before trending upward at year's end. At no point, however, do they drop below 20% and are mostly 25% or greater, signaling their ongoing relevance to discussion. Lastly, AQI Monitoring posts trend upward throughout the year, and by year's end comprise over half of all posts; the only exception is a dip surrounding June 6-13.[11]

Finally, we examine our control series, and also include our dependent variable of the censorship rate in Figure 2:[12]

Figure 2: Key Control Variables



The graph above shows a negative quadratic trend in the estimated censorship rate. Censorship is *relatively* low early in the year, relatively high in the middle (except for a dip immediately surrounding June 6-13, or days 150-200) and declines again near the end of the year. Both figures are polynomial smoothed, obscuring many short-term fluctuations that become important in regression analysis.[13] Yet censorship's overall trend is not

[11] We omitted *Jiankang* from this graph because its proportion was relatively low throughout the year; we cannot infer much from observing it visually but it might still exhibit enough variation to matter in statistical analysis.

[12] This graph shows an *estimate* of the censorship rate subject to some key assumptions (see Appendix B). While we are confident that the true rate is well above zero, it is possibly somewhat lower than the .6-.8 range shown here over much of the year. For our purposes, however, the rate's level is of less importance than its relative fluctuations over time, and these show a clear pattern robust to a wide range of assumptions.

[13] To be clear, we use the original, not smoothed series in statistical modeling.

inconsistent with our theoretical predictions; it begins low, climbs leading up to June 6-13, noticeably dips around these dates, rebounds, and then declines toward year's end, indicating the state's effort to reassert control after June 13, but also potentially the presence of some tolerated speech.

The other control series, News, loosely tracks AQI Monitoring and trends upward throughout the year in contrast to *Political* and *Physical Harm*, which move in the opposite direction. Since both News and AQI Monitoring represent information flows more amenable to (or even influenced by) the state, the overall pattern across both figures is consistent with our story of a Chinese state on the defensive prior to June, strategically reactive during June 6-13, and pursuing a more proactive mixed strategy thereafter. Finally, Figure 2 speaks to an auxiliary question: the impact of actual pollution levels on both sentiment categories, and the censorship rate. To view the AQI Index alongside the proportion measures, we graph it as a ratio of the AQI scale (which ranges from 0 to 500) over a value of 429, which was the highest reading recorded during 2012 and considered "Hazardous" to human health.[14] Using this ratio, we find that pollution spikes in January, declines during summer, and increases again in the fall. One reason that the estimated censorship rate is low at the very beginning of the year (January) is beacuse air pollution was visibly bad, which is something that would be difficult to cover up through censorship. This is a case where harsh censorship could backfire, since *visible censorship cost* would be high.

Overall, the summary statistics and graphs show a general difference in category proportions between the Static and Adaptive phases. To go beyond these descriptive statistics, next we model the relationship between these proportions and the censorship rate.

[14] We take the un-smoothed plot of this ratio to make pollution's rapid short-term fluctuations more apparent.

5.1 Generalized linear models (GLM) with autocorrelated error correction

In this section, we consider the statistical relationships between our sentiment measures and the censorship rate. To do this, we compare regression models for January 2 - June 5 with those for June 14 - December 30, the periods before and after the June episides that we argue shifted censorship policy. While we do not make specific predictions for each coefficient, in general we expect the directions of significant effects to resonate with Tables 1 and 2: *Political* measures should positively correlate with increased censorship during the Static Phase, *Physical Harm* measures should show weak or no relation, and *Scientific* measures should be weakly negatively correlated. Measure signs should diverge in the Adaptive Phase with *Political* measures positively, *Physical Harm* measures negatively and *Scientific* measures strongly negatively correlated with increased censorship.

Since our measures consist of time series, we cannot use a standard linear model like OLS because the assumption of error term independence across observations is likely violated. A second issue is that our dependent variable is a proportion, while OLS and other models assume the dependent variable can take on any real number. To address these problems, 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 then deal with autocorrelation by employing Newey-West standard errors. Newey-West models require specifying the model's maximum lag order, for which we rely on the Akaike Information Criterion (AIC). Based on the AIC results, we chose a lag order of four.

We incorporated this information into our model in two ways. First, we included the *observed measures* of lags 0-4 for all independent and control variables, and included the censorship rate's own lags 1-4 on the right-hand side. Second, we set the error term maximum lag at 4 – this should capture any residual interdependence among the series not accounted for by the included variables. Taking first the Static Phase, Table 4 shows average marginal effects.

Table 4: Average Marginal Effects for Static Phase

DV: Cens. Rate	Model I	Model II	Model III	Model IV
L.Cens. Rate	0.269***	0.275***	0.281***	0.288***
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
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*
News		-0.043	-0.031	-0.064
L.News		0.126*	0.094	0.068
AQI Index			0.099*	0.097*
L.AQI Index			-0.103*	-0.123**
Health				0.072
L.Health				0.005
Jiankang				0.379*
L.Jiankang				0.066

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

The table presents four model specifications and displays lags zero and one.[15] Model I consists only of the key independent measures for *Political* and *Scientific*; the *Physical Harm* measures are absent from the baseline model because we found that with one exception, none of them were significantly related to the censorship rate. Nonetheless, we kept them in the analysis in Model IV, as they might still be (weakly) correlated with our dependent and independent variables. Model II adds News, and Model III further adds the AQI Index. Looking at the results, we immediately note that lag one of the censorship rate is positive, significant and large. Given our prior understanding of censorship as typically reactive with some lag to sudden bursts of online controversy, we were not surprised that it was autoregressive. Periods of increased censorship following breaking incidents typically last for a few days: censors usually delete the majority of targeted content shortly after an incident, then keep censorship high over subsequent

[15] Although the actual regressions were run with lags two through four also included, and the coefficients in Tables 4 and 5 reflect this influence, we omit reporting these results for brevity's sake and because we are interested only in more recent lags' effect on censorship.

days.

Turning to the key explanatory variables, Anti-Government lag one is positive and significant in all models, while U.S. Embassy lag one is positive and large but only significant in Model III. Given the significance of *Jiankang* – our one exception to overall null findings for *Physical Harm* – alongside the lack of significance for U.S. Embassy in Model IV, we suspected that both keywords were closely related and frequently appeared together in a single, recurring post. In fact, a brief look at the post data revealed that the two keywords did appear together fairly often; out of 2114 posts through June 5 containing "*jiankang*" and 1020 posts containing the word "*shiguan*" ("embassy"), there were 268 posts that contained both keywords. 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 especially sensitive. Although our findings regarding *Physical Harm* overall are null, this observation does support a more nuanced claim that even health-related posts can trigger higher censorship when linked to *Political* content.

Next, regarding the *Political* measures, our overall takeaway is that increases in these sentiments did lead to increased censorship with a lag of roughly one day. However, the fact that Anti-Government is consistently significant and moderately large while U.S. Embassy is not suggests that the two measures, while both sensitive, really represent divergent sub-categories within *Political*; indeed, the correlation between the two is (-.249, $p < .01$). One reason may be that U.S. Embassy proxied for a more heterogeneous collection of Weibo posts than our anti-government measure which more narrowly captured views critical of the regime. At any rate, the results suggest that at least before June 6, the government may have not viewed posts mentioning the U.S. Embassy dispute, even if negative, as threateningly as those more explicitly critical messages of Chinese leadership.

The other key results for Table 4 concern AQI Monitoring and PM 2.5, which are consistently negative and significant across lag one. Across both phases, we found that PM 2.5 was a consistently significant predictor; indeed, it represents the best measure we have with respect to capturing the *Scientific* sentiment. Together, these two results suggest three points: first, that the censors clearly differentiated 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 was censored less; and third, that AQI monitoring data *overall* predicted reduced censorship despite its frequent co-occurrence with keywords that predicted the opposite, suggesting that censors may have distinguished between air monitoring reports from Chinese sources versus the U.S. Embassy.

Table 5: Average Marginal Effects for Adaptive Phase

DV: Cens. Rate	Model I	Model II	Model III	Model IV
L.Cens. Rate	0.515***	0.462***	0.444***	0.417***
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
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
News		0.322***	0.318***	0.312***
L.News		-0.165**	-0.178**	-0.166**
AQI Index			0.036	0.031
L.AQI Index			-0.034	-0.041
Health				0.033
L.Health				0.043
Jiankang				0.056
L.Jiankang				0.019

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

We can now compare the results for the Static Phase to those for the Adaptive Phase in Table 5. As with Table 4, censorship is autoregressive, and here its first lag has an even stronger effect. U.S. Embassy is now positive, mostly significant, and much larger than before, with estimates ranging from .438 to .567. This shows a clear distinction with the Static Phase, and we interpret it as the government's clear intent to shut down

Embassy-related discussion after June 13. As further support, 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 is unsurprising since we expect direct criticism of the government to always lower its *credibility payoff* to not censoring. However, generic government criticism tends to be less sensitive than posts linked to a specific incident or event, since the latter has greater potential to catalyze online collective action (King, Pan and Roberts 2013; 2014). Therefore, the relatively small size of this effect makes sense. Finally, Domestic vis-a-vis Foreign is now positive and significant in Models I-III. We interpret this similarly to the other two *Political* variables as evidence of the state's determined effort to silence critical discussion after June 13, even domestic-foreign comparisons not otherwise criticizing authorities.

A third key finding for Table 5 are the coefficients for AQI Monitoring and PM 2.5, which for lag zero are negative, highly significant, and large. The fact that these results obtain for lag zero is also meaningful, since they suggest an immediate and strong relationship between surges in PM 2.5 discussion and relatively lower censorship. While due to potential endogeneity we cannot claim with certainty that increased PM 2.5-related speech caused reduced censorship, we can assert that at the very least, surges in PM 2.5 discussion from June 14 onward did not lead to *increased* censorship. Government officials and censors do not appear to have considered PM 2.5 talk threatening; indeed, it is possible that they encouraged it. With regard to AQI Monitoring, since much air monitoring data now comes from local governments, the predominance of this data on certain dates leading to lower censorship makes sense, suggesting that especially during times when local governments were successful in broadcasting more monitoring data into *Weibo*, authorities viewed the online environment as less volatile or even wanted to promote the sharing of government released data to show the government's responsiveness to public demands. These results support our claim in Table 5 of a divergence between how the government censored *Scientific* versus *Political* sentiments after June 13. Finally, in contrast to Table 4, *all Physical Harm* measures are insignificant. We interpret this to mean that these measures did not overlap with *Political* as they did earlier.

6 Conclusion

In conclusion, we find that our theoretical predictions from Tables 1-2 are generally well illustrated by the data. We observe strong statistical evidence of a difference between the January - June, and June - December periods in the degree and speed with which the *Political* sentiment category triggered increased censorship versus the *Scientific* category, particularly our PM 2.5 measure. Concerning the *Political* category, while we do find some limited support for higher censorship of the U.S. Embassy keyword, especially in the Static Phase, the evidence is strongest that Anti-Government comments were the most likely to trigger rapid censorship. That said, such a response by the censors was strongest in the Adaptive Phase.

On the other hand, the *Scientific* category, especially PM 2.5, consistently predicted *reduced* censorship, with the effect stronger and more rapid in this latter phase. Due perhaps in part to the state's own elevation of scientific standards as guiding policy (Fewsmith, 2004), as well as leaders' efforts to address pollution that were already underway as of 2012, they were likely more inclined to tolerate even critical speech so long as the focus remained on PM 2.5 data rather than broader anti-government criticisms or domestic-foreign comparisons. In this sense, by allowing the PM 2.5 keyword, leaders were able to signal to *Weibo* users the concept's acceptability in official state discourse and thereby gain credibility with these citizens. Conversely, by briefly allowing more critical speech to go relatively uncensored during June 13 but censoring it harshly thereafter, leaders acknowledged public anger during its peak while signaling that they would not tolerate ongoing dissent.

Although our main results were consistent with our predictions, there were a few surprises. First, we were somewhat surprised to find that censors did not appear to give any special treatment to posts about pollution-related *Physical Harm* except insofar as these posts overlapped with *Political*. Second, in some specifications the News variable positively predicted censorship. Further research is needed into why news content on *Weibo* would provoke censors. Third and finally, the U.S. Embassy AQI Index in Beijing predicted increased censorship during January-June but not June-December. While we did not specifically foresee this result and so cannot consider it validation of our argument, the divergence across time periods is generally consistent with the idea that censors sought to limit even discussion driven by actual air pollution levels before June 5 while allowing it after June 13. While a full analysis of the AQI Index's interrelationship with each sentiment time series is beyond this paper's scope, our tentative analysis does find that real-world pollution indeed matters for online speech and censorship.

Overall, the results support the idea that China's leaders have the sophistication (and capability) to selectively censor social media in a pattern that seeks to maximize appearing responsive to the demands of social media-using demographics for clean air and quality of life, while minimizing sentiments that make the Party-state or leaders look vulnerable and weak. Allowing scientific and health-based discussion of air pollution, while carrying some political risk, is not nearly as risky as permitting comments that directly politicize the issue and frame it in terms of the state's systemic inadequacy. However, such a finding requires several caveats. We are *not* claiming that leaders are able to micro-manage either the bureaucrats (such as the recently-created Cyberspace Administration of China, or CAC) or the in-house Internet company censors that actually oversee post deletions and keyword blocking. Nor do we maintain that leaders can foresee the direction online sentiment will take and "steer" it in real time. Beyond this, the plausibility of *responsiveness benefit* does not even require that leaders consciously strategize as perfectly rational agents. In reality, such strategic shaping of online discussion, where it exists, is likely partial, sub-conscious and developed through trial and error, what former Paramount Leader Deng Xiaoping called "crossing the river by feeling the stones". Thus, *responsiveness benefit* and *image harm* (along with the other factors) are not a theory of leaders' censoring behavior but rather a framework to understand the incentives and trade-offs they face.

With that in mind, it is reasonable to infer that for incidents like the U.S. Embassy dispute that evolve over several months, leaders at some point will be able to issue fine-grained orders that selectively filter online discussion on sensitive issues. And during moments of crisis that grab their attention either due to external influences, or as unintended consequences of Party officials' own doing – such as officials' June 6 and 13 statements – China's elites are capable of rapid and decisive interventions.[16] Yet the act of intervening is itself not costless, especially during moments of heightened public awareness. While it certainly does not conclusively support the existence of *visible censorship cost*, the fact that censorship puzzlingly dropped on June 13, a date that witnessed both a large volume of *Weibo* comments and much anti-government speech, suggests that leaders may be aware of the potential "backlash" cost from censoring at such times, as netizens may infer that the state is trying to cover up bad news.

6.1 Broader implications and future research

Overall, the paper shows how online censorship in China can vary based on both timing and the framing of the issue. It also illustrates the importance of cost-benefit calculations in authorities' decision-making. The resulting pattern of censorship is not just blanket repression of any discussion or mention of air pollution, but rather a balance between repression, tolerance, and even encouragement of some sentiment strands over others.

Aside from illuminating the state's censorship strategy, the results also show a major growth in Chinese public awareness of air pollution's harmful effects, the extent to which citizens expect their government to address the problem, and the government's response to this increasing public pressure. While such awareness is rooted in numerous factors such as higher education levels and increasing citizen emphasis on quality of life, *Weibo* itself has arguably played a role in catalyzing this awareness, and potentially in accelerating changes in government policies on air pollution reporting and remediation. In our data, real estate mogul and outspoken blogger Pan Shiyi was very active during June 2012 in calling on the government to be more transparent with PM 2.5 data, and some of his posts were very widely re-tweeted around June 13. 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 decisions to release air pollution data with

[16] Fieldwork by one of the authors examines how such rapid interventions are bureaucratically possible. See first author (2016).

the PM 2.5 measure.[17]

In response to the events of 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 (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 Chinese Premier Li Keqiang declaring a "war on pollution" at the National People's Congress in 2014 (Reuters, 2014). While our case study has focused on how censorship relates to government responsiveness to online citizens rather than actual policy change, it is possible that public pressure on *Weibo* regarding air pollution may have played a role in accelerating real-world action.

We envision two areas for future research, both related to media control under autocracy. The first continues to focus on China and considers whether our findings still hold in the changed political climate brought about by President Xi Jinping, who has presided over a crackdown on media freedom unheard of in recent decades. Many observers have argued that what appeared to be a program of "smart" or "strategic" censorship during President Hu Jintao's final years has reverted to an older logic reminiscent of the Mao Era in which the media and increasingly the Internet are expected to serve as the Party's "tongue and throat" (*houshe*).[18] In such a climate, it may be the case that top leaders do not much care about using online space to show responsiveness to netizens. Rather, they may be tightly manipulating social media, using propaganda to show the Party's dominance of online space regardless of any public discontent. While recent arrests, interrogations and online account closures of leading celebrity bloggers certainly support this possibility, before rushing to judgment it would be useful to undertake a case study of censorship immediately following the release of the *Under the Dome* documentary. If despite the conventional wisdom, *Weibo* censorship subsequent to the documentary's release follows theoretically consistent patterns, it may be that strategic censorship repre-

[17] Interview by second author with an environmentalist at a domestic ENGO in Beijing, March 10, 2016.
[18] Interview by first author with Beijing Internet media professional, December 3, 2014.

sents a deeper logic for how the Party manages social media that has survived the Hu-Xi transition.

Finally, this paper has generated testable implications beyond China for how authoritarian states manage information during popular crises. While in a narrow sense comparable cases would include those involving environmental crises or disasters and that involve foreign actors, more broadly the theory tested here could potentially hold insights for any case in which a strong central state exercises hegemonic or near-hegemonic control over Internet media, and where significant criticism coalesces on social media about some widely shared grievance. The implications of our study can also be applied to the literature on authoritarian information more broadly. Instead of explaining noncensorship in China through a "safety valve" hypothesis (Hassid, 2012), which ignores the negative impacts from unleashing public opinion, or considering more relaxed censorship as primarily an "information gathering" exercise (Lorentzen, 2015), we provide evidence that autocrats use selective censorship to communicate responsiveness to citizens. Our case thus suggests a mechanism through which public grievances can be acknowledged, addressed, and incorporated into policy change in authoritarian regimes.

References

Aday, S., Farrell, H., Lynch, M., Sides, J., Kelly, J., & Zukerman, E. (2010). Blogs and bullets: New media in contentious politics. Retrieved from United States Institute of Peace, Washington, D.C.: http://www.usip.org/sites/default/files/pw65.pdf

Aday, S., Farrell, H., Lynch, M., Sides, J., & Freelon, D. (2012). Blogs and bullets II: New media and conflict after the Arab Spring. Retrieved from United States Institute of Peace, Washington, D.C.: http://www.usip.org/sites/default/files/PW80.pdf

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

Bennett, W. L., & Segerberg, A. (2013). *The logic of connective action: Digital media and the personalization contentious politics.* Cambridge, UK: Cambridge University Press.

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

Castells, M. (2012). *Networks of outrage and hope: Social movements in the Internet age.* Cambridge, UK: Polity.

Chen, C. J. (2009). Growing social unrest and emergent protest groups in China. In H. M. Hsiao and C. Lin (Eds.), *Rise of China: Beijing's strategies and implications for the Asia-Pacific* (pp. 87-106). New York, NY: Routledge.

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. K., & Sassoon, J. (2014). State security, information, and repression: A comparison of communist Bulgaria and Ba'thist Iraq. *Journal of Cold War Studies, 16*(2), 3-31.

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.

Farrell, H. (2012). The consequences of the Internet for politics. *Annual Review of Political Science, 15*(1),35–52.

Fewsmith, J. (2004). Promoting the scientific development concept. *China Leadership Monitor*, 11, 1-10. Retrieved from Hoover Institution, Washington, D.C.: http://www.hoover.org/research/promoting-scientific-development-concept

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

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

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.

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.

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.

Howard, P., & Hussain, M. (2011). The role of digital media. *Journal of Democracy, 22*(3), 35–48.

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.

Lang, G., & Xu, Y. (2013). Anti-incinerator campaigns and the evolution of protest politics in China. *Environmental Politics, 22*(5), 832-848.

Lieberthal, K., & Oksenberg, M. (1988). *Policy making in China: Leaders, structures, and processes.* Princeton, NJ: Princeton University Press.

Lieberthal, K., & Lampton, D.M., eds. (1992). *Bureaucracy, politics, and decision-making in post-Mao China.* Berkeley, CA: University of California Press.

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

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

Lynch, M. (2011). After Egypt: The limits and promise of online challenges to the authoritarian Arab state. *Perspectives on Politics, 9*(2), 301–10.

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.

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

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.

Shadmehr, M., & Bernhardt, D. (2015). State censorship. *American Economic Journal: Microeconomics, 7*(2), 280-307. DOI: 10.1257/mic.20130221.

Shirky, C. (2011). The political power of social media. *Foreign Affairs, January/February*, 1–9.

Stockmann, D. (2013). *Media commercialization and authoritarian rule in China.* Cambridge, UK: Cambridge University Press.

Tufekci, Z., & Wilson, C. (2012). Social media and the decision to participate in political protest: Observations from Tahrir Square. *Journal of Communication, 62*(2), 363–79.

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. (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