CHINA'S WEIBO EXPERIMENT: SOCIAL MEDIA (NON-) CENSORSHIP AND AUTOCRATIC RESPONSIVENESS

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Christopher Marty Cairns

May 2017

CHINA'S WEIBO EXPERIMENT: SOCIAL MEDIA (NON-) CENSORSHIP AND AUTOCRATIC RESPONSIVENESS

Christopher Marty Cairns, Ph.D.

Cornell University 2017

Social media's role in facilitating anti-authoritarian protests has received much recent attention. Although a handful of regimes like Tunisia and Ukraine have undergone major changes, savvy autocrats elsewhere have co-opted online space with propaganda while censoring to prevent opposition. Yet in China and other cases, we sometimes observe less censorship than conventional wisdom about authoritarian information control would predict. Why do some autocrats choose to censor selectively, and how do they actually implement such fine-grained control? In this project, I argue that allowing limited online criticism can signal regime responsiveness to public demands on issues where leaders' legitimacy is at stake. I develop this logic through a focus on China. Chinese Internet industry interviews address the why and how – i.e. the elite beliefs, and bureaucratic apparatus – behind China's selective censorship since 2011. Second, social media data analysis of online incidents on Sina *Weibo* (China's Twitter) reveals that censorship is selective even within sensitive issues. The implication of these findings is that leaders' ability and willingness to fine-tune censorship may be vital to maintaining popular support (or forestalling dissent) among increasingly educated, urban, Internet-literate publics whose views are crucial to regime survival in rapidly developing authoritarian states.

BIOGRAPHICAL SKETCH

Christopher Cairns was born in Red Bank, New Jersey but grew up in Fort Collins, Colorado to parents who have always encouraged him to pursue his dreams and inspired him to care about the power of politics to shape individual destinies. At age 18 he left Colorado to begin a journey that first reached Georgetown University's School of Foreign Service, where he majored in International Politics as well as studied abroad and worked in Latin America and Indonesia. After graduating and then completing a one-year Master's degree in Human Rights at the London School of Economics, he ended up in New York City managing digital media for a major nonprofit, and was influenced by new media's ability to both connect, and persuade citizens. While in New York he also pursued a personal interest in China and Mandarin by self-studying the language in his spare time. These disparate currents – politics, human rights and free expression, China, digital media, and scholarly research – eventually merged together in his plans for graduate school and ultimately, six years in Cornell's Government department, an experience that transformed his professional life. This dissertation is the end product of that journey.

To my parents, Charles Cairns and Deborah Hamilton, and my wife, June Pan, with whom I share the road of life.

ACKNOWLEDGEMENTS

First of all I am indebted to my dissertation committee, especially my chair Andrew Mertha. Andy always knew just the right nudge to give at the right time, striking a balance between keeping me on track, and leaving me to swim on my own as every graduate student must. His insights into how to clearly articulate the dissertation's real-world import were vital in pulling me out of jargon and into publicly communicating my ideas. He also never failed to offer moral support. As my co-author, Allen Carlson took me through the whole process of creating an article, treating me as equal colleague. He also prompted me to think hard about my target audience. Peter Enns never hesitated to ask the hard questions about my measures and research design. Finally, Daniela Stockmann read my work as a Chinese media expert, provided fieldwork strategies and contacts, and most of all led by example through her own projects.

I would also like to thank my fellow graduate students at Cornell and elsewhere. Elizabeth Plantan and Manfred Elfstrom were awesome co-authors. Wendy Leutert, Isaac Kardon, Lin Fu and many other colleagues provided valuable ideas and support. For those I have neglected to mention, it is due to a lack of space or my own oversight rather than any unworthiness on their part.

Funding for this project came from the National Science Foundation, and the Cornell Einaudi Center for International Studies, East Asia Program, and Graduate School. During my year of fieldwork I received sponsorship from Peking University's School of Journalism and Communications and would like to thank my host, Professor Wang Xiuli, as well as numerous faculty and students in PKU's Center for Social Media Research. These individuals all provided key interview contacts, ideas and feedback. And of course this project would not have been possible without the participation of numerous interviewees and contacts in China,

who for reasons of confidentiality and security must remain anonymous.

Third, I am deeply indebted to my parents, without whose unceasing love and support for the past 32 years I would not be at this point. They taught me the value of hard work and courage in the face of uncertainty, and I am forever grateful.

Finally, I must single out my wife, June Pan, who has been my inseparable life companion and fellow doctoral Cornellian ever since we met in Ithaca. From late nights in the library together, to endless discussions about research hurdles, to long shared drives to the nearest major airport, I could never have imagined a more amazing partner to share this epic journey. It may be cliché to say that marriage (and graduate school!) both fundamentally transform one's life, but with June these two events have been so interwoven that one transformation is unimaginable without the other. I could not have completed this process without her.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

PREFACE

Since I first began to brainstorm this project around 2012, the political currents in China have shifted dramatically, with regulation of social media likewise undergoing rapid change. At the time, China's Internet was abuzz with contentious examples of activist microbloggers questioning state authority: the scandal of the Wenzhou train crash, a key turning point in regime perceptions of social media's power to foment opposition, had occurred the previous year, and 2012 witnessed several more large-scale events such as Politburo official Bo Xilai's downfall and a wave of Diaoyu/Senkaku islands protests. As this project documents, Chinese leaders' reaction to this livelier climate was mixed, taking steps to shore up their ability to swiftly and decisively regulate Internet companies and their services while in practice allowing substantial room for criticism.

Just as the situation looked like such partial tolerance might become a long-term norm, however, newly installed President Xi Jinping began an ideological tightening that has affected all manner of free expression in China, including social media. In the conclusion, I argue that this ideological crackdown, as dire and disheartening as it is for individual activists, civil society and human rights in China, does *not* necessarily imply that a broader logic of selective censorship is totally invalid. President Xi and his supporters appear determined to regain societal loyalty to the center's ideological precepts, particularly for Party cadres but also within state-owned enterprises, universities and other CCP-run institutions. Yet while they may have tightened overall "public square" space both on- and offline in general, selectively tolerating bottom-up popular criticism on social media during major crises may still have instrumental value even as the state harshly represses individual bloggers. Xi's shift in direction thus presents an opportunity to test the argument's validity as long-term explanation of China's information control

strategy, rather than refuting it *ipso facto*.

To be sure, testing these conjectures will not be easy, especially in a still-closed authoritarian system like China's. To study the complex interactions between state and society as mediated by ever-evolving digital technology is to aim for a moving target. Entire platforms can become politically less relevant or irrelevant overnight (as some have maintained about Sina *Weibo* since 2013), not necessarily because of what the state does but simply because users have latched onto a "hotter" platform (such as China's WeChat). State censorship and propaganda strategies rapidly morph as officials struggle to keep pace with netizen behavior. And possibilities for data collection open at a given moment – as was collection in 2012 of this project's data – may suddenly become impossible or infeasible.

These realities make scientific standards of replicability, and broad testing of scope conditions across different samples and sub-populations more goals to strive for than minimum expectations. While researchers of the Internet under autocracy may never be able to completely overcome concerns about state-sponsored truncation of data access, they should of course continue to do everything possible (subject to ethical concerns) to push the boundaries of data collection and to be transparent about potential biases. The topic of how authoritarian states are seeking to intervene domestically (and globally) to shape the digital media environment is too important to let the perfect be the enemy of the good. And as recent studies of censorship demonstrate, this research program *has* made progress in generating results sufficiently well-warranted to both provoke debate about the logic and mechanics of authoritarian media control, and to justify and lay the foundation for ongoing research. It is my hope that this project will further solidify comparative digital media research, especially censorship and propaganda, as core concerns of political science, and of interest to scholars, practitioners and the public alike.

CHAPTER 1

CHINA AND THE 'SOCIAL MEDIA SHOCK'

1.1 Introduction

The global rise of social media has provided long-repressed populations in autocratic states new tools to mobilize, but in most cases has not led to the regime change some predicted. While any given medium – microblogs like Twitter or social networks like Facebook – in isolation is clearly insufficient to bring about regime change, protesters' intensive use of these information and communication technologies (ICTs) during the Arab Spring, so-called "Color Revolutions" in Eastern Europe, and elsewhere suggests that the role of these new platforms deserves serious attention. Yet far from deterministically leading to active societies pressuring cowering autocrats, this 'social media shock' has threatened some regimes far more than others. Social media (arguably) played a role in toppling long-established dictators in Egypt and Tunisia but in only the latter case is a democratic transition underway. Protesters in Iran using Twitter challenged the status quo during the 2009 elections but were not even able to replace the incumbent leadership, yet alone the regime. And in China, surging social media use among the population has left the Communist Party's hold on power intact, and in fact may have strengthened it. Much recent work in political science has focused on the potential of new ICTs to enable protest cascades when other, more fundamental drivers of revolution are already present. This line of inquiry asks: has social media catalyzed regime change? While this society-centric question is important, a less dramatic but equally important question re-focuses attention on *state responses* to the emergence of social media and is this dissertation's focus: why have some

authoritarian states like China appeared to better harness social media than some of their peers?

Rather than focusing on social media as a society-centric force that autocrats either have sufficient capacity to repress, or not, I focus on the degree to which states are willing and able to adopt nuanced rather than brute strategies for controlling online space. The most sophisticated Internet-censoring regimes do not face a stark choice between throttling the entire Internet, or keeping it mostly open and having to rely on costly offline means to suppress dissent. Instead, savvy autocrats have developed fine-grained means to filter unwanted opposition speech while maintaining just enough openness to reap both the Internet's commercial benefits, and to serve as a mechanism for social input. While online citizens in China enjoy lively news, games, and even microblogs where controlled but occasionally raucous political discussion is sometimes allowed, their counterparts across the border in North Korea are cut off from the global Internet, except for individuals with special authorization. Meanwhile, Internet access in Egypt remains partly free, and content filtering relatively crude despite the imposition of military rule and a crackdown on the Muslim Brotherhood. One explanation for this variation is surely these states' differing constitutional, legal and normative environments as well as the percentage of Internet penetration – online space is easier to throttle if fewer people are using it. Yet even in democracies like Turkey and India with legal protection of free expression, restrictions on online content and activity have recently tightened, suggesting that regime type and differing legal structures alone cannot account for the divergence.

This dissertation pursues an alternative explanation to explain this variation in states' adaptive sophistication in Internet censorship – on a spectrum from

what appears as a coarse binary choice to deny or permit Internet access (North Korea and Egypt) to content and search filtering so sophisticated it is increasingly invisible (China). Through a study of China as a "most likely" case of what I refer to as 'selective censorship', I disaggregate the question of adaptive sophistication into four sub-questions. First, theoretically, *why* might rulers want the capability to selectively tighten and loosen online space across issues, and over time? Put differently, how, exactly, might rulers gain instrumental benefit from such fine-grained control?

Through developing a theoretical framework (in Chapter Two) that captures the trade-offs leaders face in choosing to reduce or increase censorship, this project offers an answer to this first question: rulers, particularly autocrats, face obvious incentives to censor tightly to both thwart actual collective action and preserve an image of unity and strength. Less intuitively, however, temporarily loosening control allows leaders who otherwise keep a firm hand on the Internet tiller to benefit by implicitly *signaling responsiveness* to citizen demands for certain key reforms voiced by those who use social media the most – China's wired, increasingly educated middle class. In essence, the project tells a story of how authoritarians – if they are willing to run the risk of some collective action and reputational damage – can use social media openness to satisfy or at least demobilize this increasingly restive yet crucial population segment. I argue that the crucial mechanism through which they accomplish this is the information transparency – about responsibility or blame for 'hot-button' breaking issues or scandals – that temporarily lowering censorship enables. Transparency about the central government's role in or responsibility for the problem generates shared knowledge among online citizens that the government is on the hook to take action, a logic that I elaborate in depth later on. Nonetheless, leaders must balance this *responsiveness benefit* against three other

variables: the *image harm* to leaders' reputations from allowing opposition speech online, the *collective action risk* of not blocking social media, and the *visible censorship cost* to leaders if censorship attempts are too obvious. Chapter Two also elaborates on each of these other three factors.

In a different vein, the next two sub-questions concern how state leaders actually understand and implement such an abstract logic in practice. The second sub-question concerns leaders' *subjective* and *inter-subjective interpretation* of social media's threat/opportunity structure in response to the destabilizing shock of the Arab Spring and similar protests where social media were believed to play a catalyzing role. What lessons did Chinese elites learn from the global social media surge in 2009-11 and protests in Iran, Egypt, Ukraine and elsewhere? How did the ideological lenses through which they viewed these events shape their collective assessment that the appropriate response was to accelerate efforts at making censorship more fine-grained, as opposed to the alternative of more brute digital repression?

The third sub-question concerns state *capabilities* to effect subtle censorship rather than leaders' desire to do so. What bureaucratic structures and capacities are necessary to make designing and implementing a selective censorship policy sufficiently top-down and unitary that speaking of top elites' unified strategic intentions is warranted? Chapter 3 relies on interviews with Chinese media practitioners and Internet sector employees to derive answers to both sub-questions two and three. I find that Communist Party elites' beliefs about the value of proactive (mixing selective filtering of bottom-up voices with positive propaganda) rather than only reactive (physically silencing unwanted discussion) intervention in online space, prior experience with the Internet, and belief in technological governance

solutions primed them to recognize the Arab Spring and "color revolutions" as a threat, but also convinced them that further developing a nuanced censorship strategy was the appropriate response. Additionally, I find that Chinese leaders' success in creating a unitary, top-down bureaucracy to regulate social media, as well as the presence of pre-existing agencies capable of both 'positive' and 'negative' means of information control, allowed leaders to rapidly implement more nuanced censorship after 2011.

The fourth and last sub-question concerns both how to conceptualize and measure the dependent variable of censorship in actual social media data, and how to link observed variation in censorship to the theoretical framework's variables. Can we actually observe and measure the Chinese state (via the Internet companies that it oversees) censoring social media in real time, and if so, is the pattern of censorship consistent with Chapter 2's theorized logic? Establishing selective censorship's logic (internal validity), elite thinking consistent with such a logic, and what bureaucratic structures are necessary to speak of a unitary and strategic state are all falsifiable tests for the existence of highly nuanced Internet control in China. Yet even if all these criteria hold, i.e. leaders have fine-grained control and express a desire to exercise such control with a nuanced strategy, the censorship logic they follow may simply be different from the explanation I propose. Throughout the dissertation, I consider such alternative logics and explain why each is more inconsistent with the data than an explanation rooted in these variables.

Each of the above questions calls for a different research design to attempt an answer. The second and third sub-questions require process tracing the case of contemporary China as a single observation (at the elite politics, and bureaucratic levels), while the fourth one prompts medium-to-large N analysis of within-country

variation, and at a lower level of analysis, of within-*incident* variation across different instances of politically sensitive "breaking incidents" on social media. I choose three individual *incidents*, all drawn from the year 2012, of breaking news accompanied by heated online discussion in which the decision to censor is likely to be most salient. The specific issue in question, the basic facts surrounding the incident, and macro-political conditions (such as tensions between China and its neighbors or domestic economic conditions, to give two examples) all affect leaders' cost-benefit calculations. I apply the four-variable framework mentioned above to analyze each incident's circumstances, and then employ content and statistical analysis to measure the rise and fall of specific sentiment categories or topics during the incident, where I argue that fluctuations in these categories strongly influence leaders' propensity to censor. Using time-series and sentiment analysis, Chapters 4-6 study correlations between the categories and the observed degree of social media censorship across three issue areas – air pollution, elite-level corruption (the Bo Xilai scandal), and nationalism/territorial conflict with Japan – all of concern to China's emerging wired middle class, and that therefore (to varying degrees) provide motivation for Communist Party leaders to want to signal their responsiveness to this public.

The questions raised above have implications both within and beyond political science. Within the discipline, the project's most important broader implication is for understanding the role of new media in mediating state-society relations in one-party and dominant-party contexts, especially in states undergoing disruptive economic growth and social change. Studying social media and censorship gives scholars a direct window to observe what emergent middle classes with Internet access are demanding beyond mere material security, but more importantly, potentially provides insights into why leaders feel they must respond to this group, and

how and whether they view social media as a tool to better manage this relation while keeping a lid on outright dissent. In Chapter Two, I link the project's purpose and findings more explicitly to comparative literatures on authoritarian survival, public opinion, and the incentives behind autocratic censorship and information control. Another potential contribution is to understand state-business relations in autocracies, specifically whether administrative decentralization – long a tool to promote innovation and economic growth in China – might not be politically optimal for managing the giant Internet companies that increasingly dominate the global ICT sector, a potential finding at odds with the "fragmented authoritarianism" said to characterize the Chinese Party-state (Lieberthal and Oksenberg 1988; Lieberthal and Lampton 1992). More directly, this research matters for practitioners and policymakers because the emergence of a media-consuming middle class has long been considered a pivotal factor in democratization. Therefore, studying how the state grapples with new technology in an effort to satisfy this group's demands while remaining in power, is potentially key in assessing China's and other states' democratic prospects.

The rest of this chapter both lays basic social-scientific and research design groundwork for the study, and introduces the general background of social media and Internet censorship in China that originally inspired this project. The next section answers a fundamental question: what is the appropriate unit of analysis for studying censorship? Third, I justify a focus on contemporary China as a most-likely case of selective censorship, as well as comment on salient regime type and other characteristics likely to shape whether generalizing beyond China will be successful. This section also elaborates, drawing from the China experience, on the specific technological and human properties of social media (and China's *Weibo* microblog in particular) that give initial justification to limiting the theory's scope

only to these highly social spaces. Fourth, I more formally introduce the project's research design and methods. Crucially, the empirical chapters (3-6) do not *test* the theoretical framework in Chapter Two, but rather are themselves theory-generating in that they draw on multiple data sources and methods to dig deeply into each issue/incident, in order to empirically support the abstract theorizing with concrete examples. Fifth and last, I consider alternative explanations that also appear consistent with the observed 'selective' censorship pattern.

1.2 Units of Analysis: Breaking Incidents in Issue Areas Resonant With *Weibo* Users

An initial question in this study concerns appropriate units of analysis. How to reduce the particularly ineffable digital world into tractable units? The answer is by no means self-evident: entire countries' measurable levels of Internet openness (according to the Freedom House reports, or any other demonstrable criteria);[1] the general state of censorship across different services or platforms; the presence or absence of particular censoring practices such as blocking posts; and even micro-units like individual blog posts themselves are all candidates. For purposes of this study, however, I propose that maximum analytical leverage is to be gained by subjecting the *breaking event* or *online crisis* to analysis. The political science literature is replete with contributions that emphasize crises' special role in stress-testing regime resilience.[2] In online space, such events are the major focal

[1] See Freedom House: *Freedom on the Net 2014*. https://freedomhouse.org/report/freedom-net/freedom-net-2014.

[2] See, for example, the literature on economic crises and institutional change in its rationalist, historical-institutional and ideational variants (e.g. Gasiorowski 1995; Hay 1999; van Hooren, Kaasch and Starke, 2014). Of course, the economic shocks usually the subject of this literature

points of political social media discussion. Equivalently, they are the moments in which citizens potentially acquire *common knowledge* of factors like regime durability, responsiveness, or their fellow citizens' level of discontent with the status quo. Chapter Two highlights the importance of such moments in explaining why these online attention bursts provide leaders with a valuable opportunity to signal responsiveness to citizen demands.

Before going further, this project's focus only on 'nationally resonant' online crises merits careful consideration. What exactly makes a breaking, urgent incident 'nationally resonant', and how do we define it at least partially exogenously to both the state, and social media users? In China, negative examples of political state-society interactions that do not meet this 'national' criterion abound: land seizures, small-scale water pollution, everyday labor disputes, and cases of local-level petty corruption are all unlikely to fit the bill, unless some aspect of a localized event finds a way to resonate nationally with social media users. While negative examples are somewhat clear in that such small events have limited potential to engage any of the broad state-society dynamics I describe in this project, certainly not all 'large' events qualify as comparable units either.

Of course, what resonates nationally is partially endogenous to state media itself – extensive work (McCombs and Shaw, 1972; Iyengar and Kinder, 1987) has shown how national media shape what audiences think about (agenda-setting) and can trigger what they think about at a given time point (priming). An example relevant to the present analysis is nationalist, and particularly anti-Japanese online sentiments in China, which some authors (Rozman, 2013) have argued are a result

are typically associated with much farther-reaching institutional change than is at stake in the Chinese domestic 'crises' in this project, but the idea of unanticipated shock followed by public and elite demands for policy change does closely fit the sorts of issues and reform areas I consider, and the sorts of incidents about which discussion on social media is likely to surge.

of the state's own efforts at 'patriotic education' among China's youth. The fact that a given crisis 'resonates nationally', then, is partially a product of the state's own design. If this is the case, though, then how could events like the public outpouring of emotion in mainland China in response to escalating Sino-Japanese conflict over disputed territory in the East China Sea be 'unforeseen' and thus meet the above definition? I suggest that this apparent contradiction is logical rather than empirical: even though leaders are aware (due to their own longstanding emphasis on patriotic education) that such nationalist outbursts are likely to occur in response to breaking events, they still cannot foresee the specific details of how such online episodes will manifest (how severe, how long, and how critical of leaders themselves). A contradiction exists only if we believe Chinese leaders are capable of minutely foreseeing any and all episode-specific consequences of long-term propaganda efforts. Clearly, this is not true either for nationalism, or for other domestic issues.

Whether a particular breaking incident resonates with social media users, then, is not just a product of state propaganda. Indeed, a host of other factors such as users' socioeconomic backgrounds, education, and life experiences certainly also play a role. While these tend to be skewed toward higher levels of education, income, and younger ages, social media is still far too diverse a space for any sub-group's specific situation to result in the dominance of any particular set of topics online. Therefore, the sort of incidents likely to go viral across a wide swath of users are those that resonate with the lowest common denominator: in Converse's (1962) terms, these people are members of no 'issue public'. In practice and in the context of China, this then reduces the number of issues appropriate to include in the theory to just a few: these include nationalist themes, quality of life issues like

pollution and food safety, and the perceived level of official corruption.[3]

Finally, selecting the vague term 'crisis' for the definition deserves scrutiny. I employ it as it carries an appropriate sense of urgency for state leaders, and usually the sense of a threat to either their legitimacy, or practical ability to govern effectively. In this dissertation, a 'crisis' is a sudden, (mostly) unforeseen event that potentially exposes Communist Party weakness. Regarding the unforeseen aspect, I do not mean that top leaders must have absolutely no idea that some breaking event could occur, only that the timing and specifics of what actually occurs be unforeseen. Second, concerning 'weakness', crises with potential to 'go viral' on social media need not seriously threaten the Communist Party's hold on power, or even come anywhere close to impairing their ability to govern. In the language of Keohane and Nye (1977), Chinese leaders are often 'sensitive' to crises that play out online, but not 'vulnerable'.

1.2.1 The Need for a Catalyst

Even events that meet all of the above conditions do not always 'go viral' on social media if they lack a catalyst. In chemistry, a catalyst is a substance that triggers or accelerates a reaction. In the theory here, a catalyst is otherwise unimportant, other than that one be present, and does not itself primarily drive the dynamics of viral information spread and the state's censorship response. The point of this metaphor is that catalysts are usually substitutable, and are not syn-

[3]This is not an exhaustive list, nor do I intend for the characteristics of these issues to define the universe of cases. I justify focusing on these issues as dissertation cases because I argue that they are politically 'salient' (to use Converse's definition) among large segments of the social media public. Conceivably, other issues, such as inflation or job security, could also become salient under the right circumstances.

onymous with the underlying issue they trigger. One example from China concerns the catalytic role of a 2011 high-speed rail disaster near the city of Wenzhou that involved the collision of two trains and official attempts to both suppress media coverage of the disaster, and more insidiously, to literally bury the evidence. Both the media ban and the physical cover-up failed after bloggers posted images of officials at the scene on popular domestic microblog Sina *Weibo*, leading to a massive online outcry. In this incident, the underlying issues that determined the political sensitivity of social media discussion were transportation safety, public anger over corruption that may have caused the Railways Ministry to cut corners, and broader questions about the Communist Party's technocratic narrative of order and progress. The specific disaster – two trains colliding – was necessary to trigger public discussion, but one could well imagine a range of other transportation-related disasters accomplishing the same thing. While not a primary focus of this project, dramatic turns of events like the Wenzhou disaster are usually necessary to trigger a full-blown online crisis.

1.3 Case Selection and Scope of Findings: Why China?

This project treats China as a "maximal case" for selective censorship in order to establish the plausibility of its main explanatory logic of leaders using social media to signal responsiveness. China is not totally unique in terms of the censoring tactics it uses, level of discourse repression (online and offline), or challenge of maintaining legitimacy with a growing middle class: such attributes are common to rapidly developing autocratic countries. Rather, China stands out for both the breadth of its censorship program and the enormous resources it brings to bear

in this area. The project's goal is to establish the *minimal conditions* necessary for selective censorship and to show that it operates in at least one country; if I find that China already meets this bar, then other states like Iran and Saudi Arabia may as well. If China does not, however, then other states are unlikely to, because they would lack the following three necessary conditions for inclusion in the universe of cases: 1) large and vibrant domestic Internet companies as well as a large and active social media-using population; 2) a technologically sophisticated and functionally differentiated bureaucracy; and 3) either a single, or dominant party that does not face significant electoral competition.[4]

I consider each of these in turn, beginning first with the joint criteria of domestic Internet companies and a large social media-using population. The former criterion implies the latter; in the social media age, countries with large domestic Internet companies obviously require a large domestic user base to exist. The converse, however, is not true; Indonesia in 2014 had 83.7 million Internet users, yet U.S.-based Facebook was the dominant network with a 93.9% market share of all domestic social network users, the highest of any Asia-Pacific country. In contrast, Russia in 2012 had 84 million Internet users, and the Russian-based Vkontakte was the country's most popular social network with over 50 million monthly active users in 2014. The need for large and popular domestically based social networks is critically important for the theory because my argument above — that the state, though oversight and regulation, has fine-grained control over censorship on social network sites — should not be possible for sites that are legally registered outside that state's territory. If a site is registered overseas, a government intending to

[4]The criteria for including a state in the universe of cases are *not* the same as the 'minimal conditions' for selective censorship that are this dissertation's focus. Rather, they represent a preliminary *a priori* judgment as to the types of countries where selective censorship could possibly exist. As with the project's main variables, these scope conditions are open to challenge if extra-scope examples can be found.

implement censorship measures can block it from being viewed domestically or exert political pressure on the company or host country to remove objectionable content, but it cannot order censors to do so. Regarding a large social media-using population, enough educated, middle-class individuals must take part for popular domestic services to be a forum worthy of government attention.

Second, the state must have sophisticated censorship capabilities. At a minimum, it must be able to effectively block foreign sites and social networks, monitor content in real-time on domestic networks, quickly remove objectionable text, images and other multimedia (whether through technological or human means), and have effective enforcement ability across a wide-range of *post-hoc* sanctions: closing accounts, disciplining Internet company executives, up to questioning or jailing bloggers. All four are highly useful if not necessary for the state to be able to shape which 'hot topics' end up in front of social network users' eyes. The theory does not require that states constantly exercise control in all four areas, only that they have the capability to do so, and have a strong overall interest in fairly restrictive Internet censorship. This should hold true regardless of whether state agencies directly implement censorship, or if domestic Internet companies are tasked with doing so.

Finally, decisions concerning censorship policy should be in the hands of a single party, or dominant party that does not face meaningful electoral competition, i.e. where the electoral outcome is never in doubt. Although the theory lends itself particularly well to one-party Leninist states due to these regimes' emphasis on media as "tongue and throat" of the Party, this is not an absolute requirement so long as the party has the ability to use media and Internet control to maintain hegemony, without substantial organized opposition. The essential link here is that

in the absence of significant opposition or elections, a ruling party faces a problem in convincing newly empowered urban citizens that it "gets" their demands and will carry them out, since unlike in truly competitive authoritarian systems, it cannot commit (even theoretically) to yielding power or punishing itself.[5] This is particularly true in (relatively) resource-poor one-party states like China that both require substantial rents to ensure the loyalty of large numbers of supporters, and must rely on economic growth rather than resource extraction to do so.

To foster economic growth, such regimes must provide some amount of education and public goods, and allow the population a certain degree of freedom to engage in productive activity. In one-party states with large populations, the option of repression is also lessened by the high cost of maintaining military and security services on the necessary scale and the difficulty in ensuring these agents' loyalty to the central leadership. Thus, ideology and propaganda (and the media control needed to disseminate them) take on additional importance in such states. The rise of an urban middle class due to economic growth, however, may pose a challenge if traditional propaganda no longer resonates with this group and if wealth generation through economic growth is no longer enough, a scenario that tests the ruling party's ability to adapt to both these individuals' rising expectations for a cleaner environment, less corrupt government and other goods, and their changing media consumption habits.

This scenario, potentially present in current and former one-party states in Eastern Europe, Cuba, Vietnam and elsewhere, is precisely what the theory of selective censorship is meant to speak to, and reflects common challenges that face

[5]Personalist dictatorships are outside the theory's scope because unless they build effective parties, they face a narrow selectorate (Bueno de Mesquita *et al*, 2003) and thus are relatively unconcerned with maintaining favorable public opinion, typically using some combination of natural resources, personal networks, repression, or 'personality cult' status to remain in power.

a range of current and former one-party states as economic openness supplants the governing party's longstanding ideology. While the theory contains certain elements that may be useful to explain media control in other regime types, one-party states share an emphasis on ideology and propaganda as well as a common historical trajectory that make them an especially good fit. In the conclusion (Chapter Seven), I survey real-world examples of such states and briefly compare them to the Chinese case, although rigorous comparative analysis must await future work.

1.3.1 Why Social Media? Justifying the Analysis of Sina *Weibo*

The theoretical framework developed in this project is primarily meant to be applicable to the social media era (2009 onward). That said, given the general-sounding nature of terms such as *responsiveness benefit*, one might ask why the framework does not also apply to Internet media pre-2009, or even to Chinese media generally. In other words, what is unique about social media that enables the framework to operate most saliently within this technological form? In essence, the question speaks to how the technological (and human) characteristics of social media may limit its applicability. Before going further, it is important to clarify that this project does *not* aim to explain why the recent surge in social media globally (or in China) occurred, or whether this has causally affected the openness of public discourse. Absent a viable research design to explore the counterfactual of social media non-emergence, we cannot infer causality from this single, likely unidirectional historical event. We can, however, study social media's unique char-

acteristics, compare these to other Internet and non-Internet media forms, and induce how these may have delimited the range of state responses.

The reasons why social media (and microblogs in particular) are unique among Internet technologies could not be identified prior to beginning initial research on the project. Rather, they emerged inductively from an initial examination of Sina *Weibo's* properties in particular, and those of other social media sites more generally. Of course, this was a theory-generating rather than testing exercise: there is no *a priori* reason why selective censorship should apply mainly to social media, given that Chapter Two's framework is ultimately rooted in an understanding of the state's response to the mobilization potential of new ICTs broadly defined. Instead, while the framework's implications for technologically mediated state-society relations are substantial due to the specific predominance of social media as the definitive 21st Century ICT thus far, in my initial research this hypothesized scope limitation emerged nonetheless as a necessary caveat to the framework's broad scope. As initial justification for this limitation, the following pages offer a brief analysis of microblog technology – which instantiates characteristics also frequently present in other social media forms – as well as the human element: four groups of actors that together with the technology, define the nature of social media space. Because the goal is to generalize from the Chinese experience, I first discuss the case of Sina *Weibo* at length, and in the next section, specify conditions that bear on which elements of this experience are likely to generalize to other Internet spaces inside and outside of China.

1.3.2 One Technology and Four Actors: Depicting the *Weibo* Universe

Politically Salient Characteristics of Microblogs

This section focuses exclusively on microblogs' technological properties, leaving the following section to discuss the business of social media in China. The first such property is the ease with which any Internet-literate user can publish online. Microblogs like Sina *Weibo* all allow new joiners to open an account quickly and easily, without cost, and to begin publishing immediately. While older users or those with limited Internet experience may find microblog interfaces challenging to navigate, companies like Sina have invested extensively in making user control over publishing as intuitive as possible, and the user experience on social media sites has undergone multiple waves of refinement since the early 2000s. With these innovations, microblogs have become a fluid and (relatively) easy-to-use way to quickly share one's personal thoughts or reactions to trending events with others, and to "follow" (and where desired, rebroadcast) the thoughts of both close friends, and high-profile public figures.

The result of these low barriers to entry has been to make publishing on microblogs a ready option for users with something to say, but lacking resources to launch their own website or print publication. This, then, relates to another salient feature of microblogs: their brevity compared with prior online forums such as blogs and online bulletin boards (BBS). Blogs suffered from the disadvantage of being a long, involved format where high value-added composition required considerable thought. As a more demanding medium for both writers and readers,

blog viewership fell well short of the massive exposure that Sina *Weibo* and similar services afforded to their more successful contributors. To give an example, Han Han, a former race car driver and noted commenter on Chinese politics and society, had 210 million 'hits' or page views on his Sina.com blog as of September 2008, meaning that people (including repeat visitors) had cumulatively clicked into his blog 210 million times during its entire existence. In contrast, on Sina *Weibo* Han Han had 41,596,072 'followers' (*guanzhu*) as of February, 2015. As a rough estimate, Han Han appears to write about one post per day, and each post he writes appears on the user timelines of all of these more than 41 million followers. This puts Han Han's total number of 'hits' or 'views' on *Weibo* several orders of magnitude higher than his blog, despite it being one of China's most popular during the height of the blogging era around 2008.

The above example illustrates two important aspects of microblogs. First, they serve as focal points in online space for individuals with a pre-existing reputation to collect 'followers'. Second, these followers do not need to specifically browse to their favorite bloggers' pages in order to view content, but have it delivered directly to their user timelines for ready consumption. The first aspect is a result of a small number of microblog services gaining overwhelming market share. Users opt for service A over service B because their friends have already chosen service A: in economic terms, social media may approach being a natural monopoly. The second points to microblog designers' decision to funnel traffic into a feed or post aggregator, which by default displays on users' home pages when they log in.[6] This feature enables content to spread much more quickly than was the case previously,

[6]This feature is also found on Facebook and other non-microblog sites. The technology is not new — arguably it is derivative of "Really Simple Syndication" (or RSS) feeds that enabled blog readers to curate their favorite blogs into a single, frequently updated digest of new posts. However, microblog programmers took this concept to a new level in terms of *Weibo*-like sites' intuitive design and visual appeal.

as users may re-post (or to use the Twitter term for this action, "retweet") content they like that appears in their feed, which then sends it to their friends' feeds.

Due to these properties, microblogs are excellent at facilitating "information cascades", defined as a rapid diffusion of common or public knowledge whereby individuals learn that others in society share their previously hidden preferences toward a topic (Kuran 1991; 1995; Lohmann 1994; 2000). Previous research (Schelling, 1960; Patel 2013) has emphasized the importance of 'focal points', such as central public squares in large cities, which allow citizens to know where to go to join a collective action even in the absence of voiced coordination. In the Chinese case, equivalent 'public squares' for any sort of collective action, even speech acts, were tightly controlled prior to microblogs despite periodic surges in such focal points' use.[7] Part of the explanation for this may have been the spatial fragmentation, even cellularity of Chinese society (see Skinner 1964) in the pre-1949 and Mao eras, with individuals unable to communicate either outside of localized social networks, or across geographic lines. This resulted both from China's geography and lack of modern communications infrastructure, and from the feudalized nature of social relations prior to the 1949 revolution, and the continued segmentation of society into work units and communes under Mao.

Such cellularity began to quickly break down in the reform era, facilitated by migration from rural areas to coastal provinces, as well as telephones, fax machines, and later, cell phones.[8] Internet technologies such as email and message boards

[7]In a brief survey of contemporary Chinese history, one could reference the Hundred Flowers Campaign under Mao, Democracy Wall, and the Tiananmen movement as powerful but short-lived examples of popular contention where focal points (public letters, wall posters, and indeed, central squares) were temporarily allowed to flourish, but were ultimately subject to harsh repression.

[8]Technology's role in facilitating the creation of a national community is, of course, well studied (Anderson 1991). For a China-specific treatment of nation-building and communication technology see Zhou (2006).

further accelerated this integration. Yet despite many optimist accounts (and many U.S. officials' belief) that the Internet would foster the unprecedented emergence of a digital public square in China, the early domestic Web was highly fragmented into discrete user communities. For example, many early BBS belonged to universities, and were accessible only to students and faculty. Even on market-oriented social networking sites, user communities tended to fragment along interest-based lines, with numerous sites competing for user traffic and market share; unlike MySpace and later Facebook in the United States, there emerged no obvious front-runner during the Chinese Web's 'take-off' period in the early 2000s.

While it did not attract even a majority of Chinese Web users, Sina *Weibo* overcame such fragmentation to some extent, becoming the definitive platform for social and political commentary among more educated Internet users. For the first time in Chinese history, celebrities, intellectuals and other opinion leaders could reach users in real time, quickly and easily. Conversely, ordinary users were free to write comments or post other content which more influential bloggers might pick up and spread widely. And in another unprecedented development, they did not have to either receive official approval, or risk underground circulation in order to do so. This last feature, the lack of requirement of a publishing license from China's General Administration of Press and Publications (GAPP), or approval from the State Administration for Radio, Film and Television (SARFT) set the Internet apart from all other public media venues. Would-be publishers of newspapers, books, magazines, movies and TV programs have to seek initial approval, and receive ongoing oversight from one or both of these agencies.[9] At least regarding microblog posts that do not contain multimedia content, Internet content providers like Sina are subject to neither, although other agencies do supervise

[9]GAPP and SARFT merged into a single agency, the State Administration of Press, Publications, Radio, Film and Television (SAPPRFT), in 2014.

them. Critically, once a microblog site has gained government approval, bloggers themselves do not need prior approval from any state or Party entity in order to post. That the Chinese state, notorious since imperial times for supervising official publications while selectively cracking down on underground circulation, would relinquish prior approval authority to companies like Sina and individual microbloggers is a puzzle that this dissertation will strive to address.

Such lack of prior approval is characteristic of the Internet, not just microblogs. Yet the consequences of such an unrestrained atmosphere in generating a free-wheeling online public sphere reached perhaps their fullest expression on Sina *Weibo* after its 2009 launch. The factors behind the platform's popularity, of course, go beyond lack of regulation and include *Weibo's* commercial viability, design features, and Sina's heavy recruitment of celebrities and other public figures to participate. Exploring this tension between *Weibo's* inherent properties, Chinese leaders' desire that it exist at all, and their attempt to control negative effects to their interests requires analysis not only of the technology itself, but also the actors that inhabit it, a task I now undertake.

***Weibo* Users**

As a platform, *Weibo* is not representative of Chinese society overall, or even Chinese Internet users. While Sina boasts that the platform has over "500 million user accounts", this number is grossly inflated due to a large quantity of 'ghost' accounts. Many of these do not belong to real people, and have instead been established by various groups to boost organizations' or individuals' follower counts. Other accounts belong to real individuals, but have rarely or never been logged into. Excluding these two groups, *Weibo* has roughly 50 million "monthly active

users" – a common metric used in marketing to measure user engagement with a platform. This user group overall is better educated, more urban, younger, and more active online than both the average Chinese netizen, and the general population.[10]

These are the people who actively follow the so-called "Big V" (*da* V), a shorthand term for famous celebrities, intellectuals, entrepreneurs and others with a "verified" or identity-confirmed status by Sina, and comment on and re-post these high-profile individuals' posts. Since followers of the Big V tend to be highly educated and to hold urban professional jobs, they are also increasingly important for China's future, as leaders attempt to shift China's economy away from relying on exports and heavy industry toward consumption and services. These individuals are also of concern to China's rulers for another reason: their relatively low levels of support for the Party. Data from the World Values Survey 2007 and 2012 waves shows that having completed a university degree, living in a coastal urban city (Beijing, Shanghai, or Guangdong), being part of China's "post-80s" (*ba ling hou* generation, or younger than 33 at the time of the 2012 World Values Survey), and reporting frequent Internet usage all negatively predict support for the Party-state, with a majority of this narrow cross-section reporting that they "don't trust very much" or "completely do not trust" the CCP, a finding at odds with consistent majorities of other demographics who indicate trust for the Party when asked this question. These trends are more pronounced in 2012 compared with 2007.

The fact that this elite group in Chinese society bucks the general tendency toward continuing strong CCP support among other demographics is consistent

[10]According to the Sina *Weibo* Data Center's 2013 report, 90% of *Weibo* users belong to China's 'post-80s' or 'post-90s' generations (meaning they were born after 1980). 70.8% had at least a bachelor's degree (or higher). According to a 2012 report by the Center, 48.1% of users earned at least 3000 yuan/month.

with Geddes and Zaller's (1989) and Zaller's (1992) 'exposure-acceptance' model, which posits a curvilinear relationship, depending on education, between exposure to dominant-Party (or other dominant actor) messages, and internalization of this information. In the model, only individuals at the highest education levels are able to resist dominant information channels and thus report lower support, because only they have access to countervailing information sources that enable a critical approach to the regime. Recent work (Beck, Tang and Martini 2013, but see Kennedy 2009) finds that among China's regions, "megacities" such as Beijing and Shanghai, which are home to more educated populations on average, are the only regions to be negatively signed in predicting government support.[11]

Last, this demographic, which I will refer to throughout the dissertation as simply '*Weibo* users', exhibits one last politically relevant characteristic: declining consumption of newspapers and other forms of 'old' media, including television, instead preferring to read the news online or via social media channels. A brief comparison of data from the 2007 and 2012 World Values Survey waves showed an increase in those reporting having used the Internet in the past day or week (from 11% to 30%), with a more dramatic increase among those with at least a high school education (from 19.5% to 54.9%). Results diverged, however, between the general population, and high school graduates with respect to newspapers. While reports of accessing information via a newspaper daily or weekly increased in the population overall (from 23% to 31.2%), newspaper usage decreased among high school graduates (from 34.7% to 22.7%). Since these highly educated citizens are increasingly exposed to the greater diversity of information available online compared with those who get their news via newspapers or TV, they are more likely to encounter what Tong and Lei (2013) refer to as narratives of "counter

[11]The negative coefficient is large but insignificant.

hegemony" (p. 292). Such uncontrolled information (from leaders' perspective) may further reinforce *Weibo* users' negative evaluation of the regime as traditional state media sources (such as People's Daily) and alternative sources (individual journalists and the Big V) compete to influence this audience.

Sina Corporation

As the parent company of *Weibo*, Sina Corporation occupies a unique niche among China's largest Internet companies. To quote media scholar Yuezhi Zhao (1998), it is more 'between the party line and the bottom line' than many of its contemporaries – notably Tencent Corporation, to which Sina is often compared. Here, I elaborate a theoretical ideal type for the incentives that an Internet company lying in the middle of Chinese media scholar Daniela Stockmann's two-dimensional newspaper marketization/openness typology (2007, p. 272) would be likely to face. Stockmann considers the general correlation between newspapers' status as a 'commercial', 'semi-official', or 'official' publication (with 'commercial' publications relying primarily on advertising for revenue, 'official' papers depending on state subsidies, and 'semi-official' papers in-between), and available space for politically sensitive news reporting at that paper (Ibid). She finds that the more commercialized a paper is, the more space is available. This two-dimensional framework is also useful, to some extent, to understand the incentives of companies that own news-driven online spaces, including microblogs. This is particularly the case for Sina, as the company's most prominent role prior to *Weibo* was operating a news portal, Sina.net. This portal's surge in popularity after its 1999 launch (the same year that Sina was incorporated) was attributed, according to a survey by the China Internet Network Information Center, to its aggressive coverage of a

major Sino-US international incident, the NATO bombing of the Chinese Embassy in Belgrade in May that same year.[12] Additionally, Sina cooperates closely with two 'official' news outlets considered authoritative voices of the central Party-state — Xinhua News Agency, and People's Daily — and frequently reposts news coverage from both on its portal. On the other hand, Sina, like other Chinese tech companies, is listed on foreign stock markets. It joined NASDAQ in 2000 and *Weibo* was listed as a separate, wholly-owned subsidiary of Sina on the New York Stock Exchange in April, 2014.

Both Sina's history as a Party-backed official newspaper content-dependent business, and as publicly-traded company with substantial foreign investment thus influenced the concept behind *Weibo* right from its launch. The prevalence of 'official' news accounts sponsored by all levels of government and many Party organizations, as well as the large share of official news content as a percentage of all *Weibo* posts reflects the integral role of Party ('official') publications in providing the re-postable news content organizations like Sina need to fill sites like *Weibo* with current events, while the presence of celebrities among accounts with the largest numbers of followers reflects Sina's recognition that *Weibo* needs brand names to attract users, make money, and satisfy shareholders. During *Weibo's* 2009 launch, Sina employees were reportedly encouraged to reach out to celebrities and invite them to participate in the service, and most of China's leading figures have accounts. While Sina, like other microblog-owning companies, faces difficulty monetizing *Weibo*, it experienced rapid user growth and by 2011 held a 56.5% share of active microblog users, and was used by more than 2,700 media organizations.[13]

[12]This historical beginning is representative of Sina's role as purveyor of official news, since recent work (Stockmann 2007; Weiss 2013; Weiss 2014) has demonstrated that the Chinese government closely supervises and even guides news coverage during sensitive international incidents.

[13]Source: *Kyle.* iResearch. March 30, 2011.

In briefly laying out Sina Corp's position between government and market, the theoretical claim is that Sina should be expected to be more attuned to government censorship directives than other Internet companies. This is not to say that other companies can afford to ignore subtle pressures from Beijing as well as explicit directives, only that due to its dependence on official media content and history as central media agency partner, central government concerns should be strongly factored into Sina's core business model and its products to an extent not found elsewhere. In other words, Sina's business interests and those of the central government may not only not be contradictory, but a positive synergy may exist to a greater extent than for companies whose content is less tied to official media.

The 'Big V'

As previously mentioned, the term 'Big V' refers to users whose identities have been verified by Sina and who generally have large numbers of followers. A brief analysis of summary data provided by Sina reveals that in 2012, out of the top 100 *Weibo* users judged by Sina's own in-house metric to have the most "influence", according to each user's self-identified profession there were 13 'entrepreneur and business' types, 6 authors, 11 'intellectuals and public commentators', 37 'entertainers', 16 'TV hosts' and 17 individuals from other professions.[14] Also notable

[14]Source: http://data.weibo.com/summary/2012year/influence. Accessed March 6, 2015. Sina generated this "Top 100" index by assigning a composite "influence" score to each blogger, which in turn consisted of sub-scores for "reach" (*chuanboli*), "liveliness" (*huoyueli*), and "coverage" (*fugaidu*). The Sina metric did not sort individuals by profession; rather, I developed professional categories according to the following, using individuals' self-identified profession on their *Weibo* pages: 'entrepreneur and business' included CEOs, investors, and start-up founders; 'authors' included a range of literary figures including novelists, essayists, etc.; 'intellectuals and public commentators was a broad category that included known public intellectuals, one lawyer, and noted newspaper/magazine editors and opinion writers; 'entertainers' included actors/actresses, and singers; 'TV hosts' included hosts for a variety of programming, such as news shows, talk shows, competitions, the CCTV New Year's gala, etc.; finally, 'other' included a variety of public personalities, such as models, dramatists, cartoonists, Buddhist monks, one traditional Chinese

was that four of the list's top ten slots were occupied by some of China's most dynamic business people: Lee Kaifu, Ren Zhiqiang, Xue Manzi (Charles Xue), and Pan Shiyi. All four have used *Weibo* to speak out on sensitive political issues in the past, and have come under fire from authorities for doing so.

This, then, raises the question of why businesspeople like Lee Kaifu are so influential despite being fewer in number than traditional celebrities, and perhaps enjoying less widespread popularity. I theorize that the answer can be found in these entrepreneurs' willingness to engage in politically risky speech (i.e. in *what* they say) instead of in the appeal of their public personas (*who* they are). To be sure, celebrities in other categories, notably actress Yao Chen, have occasionally spoken out on political topics. Yet entrepreneurs stand out for their persistent willingness to do so, by calling attention to perceived social ills. On the one hand, some of their clout comes from their reputation as wealthy and successful producers who have created value for the economy and country – and critically, are admired for having done so apart from state patronage. Yet netizens also admire them for the degree to which they seem willing to address sensitive issues; entrepreneurs may feel that due to their independent economic position and status as role models for China's new economy that they have a license, if not obligation to speak out. Entertainers and TV hosts, on the other hand, remain dependent for their livelihood on access to media channels under state supervision – being too politically outspoken could end their careers. Thus, they have neither the space, nor necessarily the motivation to address political topics. However, entertainment celebrities can still be counted as politically influential 'Big V' if and when they do speak out. 'Big V', in political terms, is thus not a binary category (contrasted to all *Weibo* users who do not enjoy such exalted influence) but rather a spectrum,

medicine expert, and others not fitting into one of the above categories.

with figures like Pan Shiyi at one end, and only occasionally politically active public figures at the other.[15]

In Chapter 2's theoretical model, the 'Big V', exemplified by entrepreneurs, are those willing and able to speak out on *Weibo* to raise questions about state accountability (or lack thereof) on sensitive issues of particular concern to middle-class urban citizens. The emergence of such a group whom ordinary *Weibo* users viewed as credible, I argue, was necessary in order for selective censorship on the platform to serve state purposes. However, these individuals usually cannot themselves generate breaking incidents as they do not produce news; this role of course belongs to journalists and editors, whom I analyze next.

Journalists and Editors

The role of journalists in China, as occupying a middle ground between state and society and between planned economy and market, is the subject of a vast literature. Here, I focus specifically on journalists' role in the *Weibo* ecosystem, in relation to the other groups introduced above. From a reporter's point of view, what *Weibo* has done to the profession is to increase pressure on news organizations to be first with a story. The root of this pressure lies in the fact that active *Weibo* users tend to consume their news within the platform itself, rather than browsing to news site home pages or reading print editions. In other words, active *Weibo* users 'live' on *Weibo* to consume multiple types of content, forcing news organizations to go head-to-head in a single, highly networked marketplace of information. Nearly all major news organizations (including state outlets like People's Daily) maintain

[15]It is worth noting that Lee has become much less outspoken since the 2013 crackdown on political Big V. Chapter 7 addresses what space, if any, if left for these figures to speak politically online in the Xi era.

Weibo feeds, and these are nearly equal in influence to the Big V themselves, according to a brief analysis I conducted using Sina statistics.[16]

Because the Big V are themselves not news originators, they rely on news organizations to supply content about breaking stories that they then comment on. To be sure, journalists can also reach audiences directly through *Weibo*, but when Big V retweet, or better, retweet and comment on their articles, this greatly magnifies the stories' reach and potential credibility. Thus, journalists and Big V exist in synergy with one another. Without journalists, Big V have no reliable factual sources about breaking events on which to comment. Journalists, for their part, are greatly constrained by China's propaganda system in terms of their ability to 'frame' or editorialize political news. While they do occasionally 'play edge ball' (*da cabian qiu*),[17] attempting to push the limits of the politically acceptable, as members of state-supervised organizations they run a greater professional risk in doing so compared with the Big V, and so must be relatively cautious. During 2009-12, journalists were, however, sometimes willing to risk breaking factual information – being careful to avoid 'hyping' or editorializing on the story – and to post this information on *Weibo* even before publishing it on news websites. Because there was often a lag of a day or two before the Central Propaganda Department (CPD) would order news organizations to cease reporting, such information would often quickly spread on social media, allowing news organizations to reap the benefits of moving quickly, and giving the Big V ready material for commentary.

To a large extent, the censorship constraints faced by journalists, and those experienced by *Weibo* users are interconnected — during major political incidents

[16]In addition to creating a "Top 100" ranking for individuals on *Weibo*, Sina did likewise with the top 100 news organizations, using the same 'influence' metric as explained above.

[17]The term comes from the practice in ping-pong of hitting a ball as close to the edge of the table as possible, a shot that if successful is difficult for an opponent to return.

that leaders decide to restrict, we are unlikely to simultaneously observe tight censorship of news coverage and openness on *Weibo*, or vice versa. Thus, some aspects of the theory of *Weibo* censorship elaborated in this chapter could also apply generally to online news organizations in the social media age. My general contention, however, is that while what such organizations do via other channels may matter for the impact of media censorship overall, the fact that much of what they do is increasingly centered on *Weibo* means that a theory rooted in this platform's specific characteristics, and key actors, can perhaps provide the most insight into what strategic purposes the state has for such channels, and for journalists themselves post-social media revolution.

In noting these trends, my objective is not to argue that *Weibo's* technological characteristics as well as those of the above four groups of key actors alone have led the state to adopt any particular control or censorship strategy for the platform. Rather, the goal is to characterize *Weibo* as a space with unique importance for leaders' efforts to influence public opinion among a skeptical portion of the population. While the above combination of *Weibo's* technological properties, *Weibo* users' resistance to state propaganda and negative view of the regime, Sina's role as a news-disseminating company, and the Big V's and journalists' prominent social role and influence did not dictate any particular strategic response from CCP leadership, I hypothesize that because of these characteristics, leaders had especially strong incentives to pursue a selective censorship strategy here compared with through other channels. Thus, while comparable to other social media, *Weibo* represents a 'most likely' case for selective censorship among Chinese media platforms, making it a good plausibility test for the framework. In the next section, I develop conditions that state which aspects of this section's analysis are likely to generalize to other platforms and countries.

1.4 Research Design

This section introduces the project's main research design and claims. Again, the claims are not formal hypotheses to test the theoretical framework, but rather guide the process of weighing case-specific evidence to support the framework's plausibility and serve as a prelude to actual falsifiable testing in future work. The claims are of two types: necessary conditions about a unified state and bureaucracy, and probabilistic claims about expected variation in censorship both across, and within specific issues/incidents.

1.4.1 Qualitative Design: Process-tracing Elite Thinking and Bureaucratic Restructuring

The necessary conditions relate to sub-questions two and three, and fall into two groups. The first concerns what factors make leaders likely to react to the 'social media shock' by concluding that a selective censorship strategy, rather than brute repression or total opening, is the best course of action. First, I argue that elites subscribing to a *Leninist* view of media should be predisposed to support selective censorship in the social media age. The logic is that in Leninist systems media channels are supposed to serve as instruments of the ruling party not only to control, but to motivate and direct society. Such control entails fine-grained rather than coarse censorship and a composite strategy of both positive propaganda, and censorship of opposition. This is because leaders wish to allow *some* popular voices to be heard, even provocative ones, if they serve the Party-state's interests while suppressing genuine dissenters, i.e. to encourage *loyal* mass input.

A cruder strategy, for example merely banning most citizens from going online as in North Korea, would not achieve this more profound goal of fine-tuning the censorship/propaganda mix in order to deploy online citizen voices in support of regime legitimation. The presence of this Leninist indicator is clear in the Chinese case, but analogues may be found in other systems including non-Communist countries. A second claim is that if elites have *prior experience* with sophisticated Internet censorship, they are likely to also apply such an approach to social media. More formally, this claim equivalently states that selective censorship exhibits path-dependence within a country's trajectory of Internet censorship. A third claim is that a *belief in technocratic governance* predisposes elites toward selective censorship. Along this dimension, countries vary in the degree to which leaders view technology as the answer to emergent governance problems.

The second group of necessary conditions (sub-question three) concerns what bureaucratic structure and capabilities are necessary to effectively implement a varied censorship strategy. The most important of these is that selective censorship only works under a *unitary* state, i.e. a censoring bureaucracy controlled by the central government, and with high-level agencies that are directly under the very top leaders' direct supervision in charge. In other words, selective censorship, of which the Internet companies who run social media platforms are the ultimate implementers, is likely inconsistent with a fragmented or decentralized authoritarian system precisely because the concept itself entails conscious, top-down design leading to systematic variation in censorship. Second, if we define 'censorship' to include both efforts to shape what content does appear before the social media public (positive censorship) as well as what does not (negative censorship), then to be effective at selective censorship a state must have bureaucratic actors capable of doing both – i.e. it must have the ability to agenda-set online by *generating* as

well as blocking or deleting content. If the state can only censor through negative (restrictive) means and loses all ability to shape public discourse more proactively, then its cost-benefit analysis is likely to skew strongly in favor of blanket censorship, since uncontrolled online discussion risks a rapid cascade of anti-regime mobilization or at least seriously damaging leaders' reputations.

The above two groups of conditions are highly comparative in nature and might appear empirically intractable if we only consider the Chinese case. To be sure, cross-national comparative work is a vital next step in the research agenda this project represents. Yet empirical leverage can still be gained from only studying China by process-tracing developments in leaders' thinking regarding Internet censorship, and their efforts to re-shape the bureaucracy in ways that might approximate the unitary state ideal. A single case contains many potential 'within-case' observations (see King, Keohane and Verba 1994; Brady and Collier 2004). To be more precise than the usual understanding of a 'necessary' condition, I argue that the claims regarding elites' inter-subjectively shared understandings of social media constitute "INUS" conditions (Mackie, 1974).[18] Chinese elites' Leninist view of media, their prior experience with Internet censorship, and their belief in technology were all individually insufficient (but necessary) but jointly sufficient to lead them to selective censorship as the solution. However, such a chain was unnecessary to bring about an appropriate ideational basis for selective censorship, as one could imagine other constellations of elite thinking in other countries that might produce the same result. One implication for future research, then, is that while tech-savvy systems with residual post-Leninist influence (like post-Soviet Russia) are a good place to begin comparative analysis, attention should also be paid to other media ideologies and attitudes toward technology that might also

[18] According to Mackie, an "INUS" condition is an individually necessary (non-redundant) part of a larger sufficient but unnecessary condition that enables some outcome.

ideologically ground a selective censorship strategy.

In contrast to elite thinking as INUS condition, my understanding of bureaucratic capacity is simpler: *all* related conditions in this area are necessary, and collectively they are the only possible means through which selective censorship may be implemented. In other words, selective censorship should not exist in any country if a) the state is non-unitary regarding censorship, or b) the state does not have both positive (agenda-setting) and negative (suppressive) means to shape what content appears in front of social media users. This stringent standard reflects what I argue is the critical importance of a unified, disciplined state for autocrats wishing to take full advantage of social media, and helps direct empirical analysis.

To actually carry out this process tracing, I rely on interviews conducted over seven months in China with journalists, academics, and Internet company employees. Short of top-level elite interviews, these are some of the best accessible sources to give insight into top Communist Party member thinking regarding social media. Chapter Three gives more specifics, and presents interview evidence by weaving a narrative of elite thinking and efforts at bureaucratic reform immediately following the events of 2011. In Chapter Three I also draw initial, descriptive inferences taken straight from this evidence. In the conclusion (Chapter Seven), I then summarize the evidence in light of the above claims.

1.4.2 Quantitative Design: Predicting Breaking Event-level Variation in Censorship

The final set of claims follows a standard quantitative approach for empirical analysis: to state predictions of outcome variation based on theory and then to test these probabilistically in medium to large-N data. As stated above, this project's main unit of analysis is breaking, nationally-resonant crisis incidents: censorship should be expected to be high, medium or low for any given incident. However, at a lower analysis level I also expect censorship to vary *within* incidents across two dimensions: over time, and across *issue-frames*; Chapter Two gives detailed predictions as to what pattern this variation is expected to follow.

The medium and large-N variation in the research design thus primarily occurs within each unit or case. Across whole incidents, I adopt a qualitative "case-control" approach to create an overall expectation regarding censorship for each incident according to the issue area it belongs to, as well as its specific circumstances. The three incidents chosen, all from activity on Sina *Weibo*, are discussion of Beijing air pollution in 2012 (a 'least likely' case of *high* censorship), the downfall of Chinese politician and Politburo member Bo Xilai (a 'most likely' case), and the August and September, 2012 flare-up in Sino-Japanese relations over the disputed Diaoyu/Senkaku islands (a 'moderately likely' or 'mixed' case). Each issue/incident further emphasizes a different variable from the four variables that together comprise the theoretical framework. For example, *responsiveness benefit* is most salient in the air pollution case, while the threat of collective action is clearly present in the Diaoyu case.

Then, within each incident, I use two techniques – sentiment analysis and

time-series statistics – to draw inferences. Sentiment analysis allows for the disaggregation of all topic-relevant speech on *Weibo* into discrete categories, many of which represent issue frames that if allowed to propagate, would yield greater benefit (or risk) to the state. Using unique measurements of social media censorship in real-time from data gathered by University of Hong Kong researchers, I can then measure the censorship rate across different frames to see if observed high or low censorship matches Chapter Two's predictions. Similarly, time series analysis of fluctuations in the censorship rate over the course of multi-day or multi-week incidents (such as the 2012 Diaoyu/Senkaku dispute) allows for testing predictions of high/low censorship based on changes in factors that affect political sensitivity from the leadership's perspective, such as street demonstrations. The exact statistical models and sentiment analysis techniques used vary slightly across Chapters 4-6 due to characteristics of each incident, such as incident duration and the number of sentiment categories present, although I am attentive to the consequences of such differences for the comparability of findings across chapters.

Before moving on to the dissertation's main theoretical argument in Chapter Two, the next and last section considers the possibility that variation in social media censorship in China and elsewhere might appear systematic even if one or more of the above claims are incorrect – in other words, the possibility that alternative pathways of shock and elite reaction could have led to verisimilar observed outcomes.

1.5 Alternative Explanations

In this section I consider alternative explanations that could account for ob-

served variation in censorship across a range of *Weibo* 'hot topics' in different issue areas. These explanations fall into one of three categories: apparent occurrences of selective censorship despite the 'unitary state' claim being false, 'rational' explanations that differ from the theory in Chapter Two, and non-rational explanations.

1.5.1 Alternatives to a 'Unitary State'

First, there are a number of plausible scenarios under which the framework's simplification of the state as a 'unitary' actor may be inadequate. These scenarios are drawn from the Chinese experience to fit this study's main inferential goal of establishing China as a case of selective censorship, but may find resonance elsewhere to greater or lesser degrees. One such possibility with much support in the China politics literature is the 'bureaucratic politics' model (Lieberthal and Oksenberg, 1988; 1992). In this model, policy outcomes are the result of bargaining between different ministries and levels of government with a regulatory stake in a particular area. To give a hypothetical example, one could imagine the Communist Party's powerful Central Propaganda Department (CPD), its agency in charge of information technology (the MIIT), and the State Administration of Press, Publications, Radio, Film and Television (SAPPRFT), which is nominally in charge of regulating all types of video, competing over jurisdiction to enforce censorship of "Internet content".

The bureaucratic politics model is considered foundational by most China politics scholars, and should not be dismissed lightly. In Chapter Three, I draw on interviews and analysis of publicly available documents to argue that while multiple agencies have indeed engaged in 'turf wars' over the right to regulate Internet

content since the late 1990s, by around 2011 a clear hierarchy had emerged with the newly-upgraded State Internet Information Office (SIIO) – an agency that reports directly to the State Council (China's highest executive body) – at the top, and an increasingly clear division of labor among subordinate ministries. Some inter-bureaucratic conflicts do exist, of course, and have affected the censorship process, but I show evidence that such clashes have not fundamentally impeded top leaders' ability to implement the censorship policy they want, and to quickly implement decisions on 'hot topics'.

A second possibility concerns divisions or lack of responsiveness to breaking incidents among elites themselves. According to this argument, elite interests are not sufficiently harmonized, or elite attention ample or focused enough regarding whether and how much to censor across a range of incidents, for them to give clear orders (or set sufficiently robust guidelines) for subordinates who actually interpret these orders for Internet companies. Thus, subordinates are unsure of what to do during breaking incidents and so defer up the chain, leading top officials to either hammer out consensus, or await orders from those at the very pinnacle of power, with either possibility leading to a delay of days during which nothing gets censored. This could be avoided if elites could agree on a comprehensive-enough 'rules-based system' for what to censor. In practice, such a system is likely to be unworkable given the variety in what online breaking incidents are deemed 'political', or 'sensitive'. Chapter Three also addresses this possibility empirically, showing that while elites' responsiveness to breaking incidents and differences of opinion are certainly problems, they have nonetheless been able to agree on general 'guidelines' for Internet censorship, and to intervene quickly in online incidents deemed of national importance.

A third possible explanation has some evidence to support it, namely that various forms of corruption, such as in-company censors accepting money to delete *Weibo* posts, explain much variation in censorship. Under such an argument, powerful individuals have enough money and connections inside the censorship bureaucracy to get certain posts scrubbed from the Internet. This phenomenon exists, and has recently been empirically documented.[19] However, while such practices likely did affect censorship decisions regarding small-scale, localized scandals, especially in the final years of President Hu Jintao's administration, they are least likely to matter in the sort of high-profile breaking events that fall under the theory's scope. The reason is simply that when broader strategic concerns are at stake, top leaders are unlikely to allow petty corruption to influence such an important decision.

1.5.2 Elite Rationales for Selective Censorship Other Than *Responsiveness Benefit*

A second category of explanations treats state behavior as unified and rational in the sense that leaders decide whether to censor based on an instrumental logic of promoting their own survival. However, besides using non-censorship to signal responsiveness, multiple competing logics are available in the literature. One major alternative is the 'venting' or 'safety valve' metaphor advanced by Jonathan Hassid (2012). In this account, Chinese leaders refrain from censoring when netizens get so fired up about an issue that their anger could pose a threat to social stability if not released. In this explanation, newspapers play a key role in covering a sensitive

[19]Source: Xinhua. September 3, 2012. "Paid Posts Deletion' Reveals Internet Corruption by Outsiders."

issue, thus providing bloggers voice and space to 'blow off steam'. A critical facet of the 'safety valve' metaphor is then that the state, via control over print media, remains in control of the 'venting' process, just as a well-designed safety valve could dissipate the pressure of excess steam without breaking.

While there may be some situations where the metaphor is apt, it ignores the role of information cascades and common knowledge diffusion during incidents that resonate deeply with the social media-using public. Just because newspapers are under control and provide the seed for microblog commentary does not mean that bloggers will not then spread viral knowledge of their own about a particular framing of events that deviates from top leaders' wishes. In metaphorical terms, I argue that the distinction between 'safety valve' and Hassid's other metaphor of blogs as 'pressure cooker' is less than the author suggests: safety valves can also explode. Once the state opens the valve, it has no guarantee that publicly expressed discontent will not snowball and turn more vehemently against leaders, rather than burning itself out. The point is then not to entirely dismiss the logic of allowing 'venting', but rather to note that it carries non-negligible costs, and is far from an automatic solution whenever netizen discontent arises. This, then, suggests a more purposive rationality in which leaders weigh the cost of allowing negative speech to spread against other strategic benefits.

Another 'rational' explanation is the 'information-gathering' benefit proposed by Peter Lorentzen (2014) where the state allows some Internet openness in order to learn the sources of popular discontent, without allowing too much society-wide knowledge of such discontent to develop. This explanation has great value and plausibility for theorizing a major motivation behind central government investment in a vibrant Internet sector – it provides an unprecedented, low-cost means of

surveillance. While this might explain why some political discussion online is tolerated, however, it cannot predict specific variation in censorship case-by-case, for two reasons. First, provided that political speech online exists at all, Beijing can censor to a great extent and still collect useful information. While I lack conclusive evidence, it is widely believed (and some of the field interviews cited in Chapter Three suggest) that China's major Internet companies regularly share data with various government agencies. Both pre- and post-censorship data are capable of being shared in this way. In other words, high censorship and information-gathering about population discontent are not incompatible goals.

Second, this explanation does not work well for viral issues of nation-wide interest to social media users. If censors' goal in controlling any medium, including the Internet, were to prevent common knowledge diffusion while merely gathering information, they would have little reason to allow any sensitive topics on social media that carried collective action risk or could potentially harm the state's image – a logic at odds with the variation observed during *Weibo's* height in 2011-12. The state enjoys a range of online channels (including websites specifically set up for citizens to report corruption or other problems), and certainly need not open the digital floodgates to achieve its information-gathering objective. The information-gathering hypothesis makes sense for the Internet as a whole (and possibly for more mundane events and everyday social grievances even on social media), but it cannot account for dramatic instances of non-censorship that draw the nationwide attention of all social media users.

1.5.3 Non-Rational Explanations

A final category of explanations addresses the possibility that other factors besides instrumental rationality determine variation in social media censorship. Anticipating a possible objection, I should clarify that invoking the role of leaders' ideas and beliefs at this level of variation is *not* the same as saying, as I do earlier, that ideas such as Leninist media theory mattered in how China's leaders interpreted the social media shock and shaped their response toward pursuing nuanced censorship. Instead, here the focus is on how leaders' beliefs and values may matter in explaining variation in their decision to censor across breaking events *after* they have agreed that the ability to exercise such fine-grained control is a good thing, and have put in place the bureaucratic infrastructure to accomplish this.

It is only at this micro level of analysis, then, where I argue that leaders are strictly rational. Nonetheless, the counter-argument that ideas and beliefs matter even at the tactical level of micro censorship decisions is compelling and deserves consideration: China's leaders may decide to censor based on whether they think particular online grievances are legitimate, with these perceptions being rooted in leaders' shared sense of identity with the online masses: what they think the Chinese people overall have a 'right' to expect, and therefore can acceptably demand that their leaders deliver. In Chinese history, expectations that the government will deliver material security, and defend the nation's interests and honor against foreigners have traditionally fallen into this realm. These explanations argue that one must analyze elites' broader social context, ideology, etc. before any means-ends analysis becomes useful. Those relevant to predicting social media censorship come in two variants — individual, and collective. First, an individual-level, belief-driven account would contend that differing beliefs between individual lead-

ers determine what gets censored and how much. The most likely version of such an argument would be that the relative openness on *Weibo* in 2011-12 was due to Hu Jintao's and possibly other Politburo Standing Committee (PBSC) members' general openness to 'public opinion supervision'; this small, unique set of individuals believed that prominent netizens and 'Big V' could play a role in holding the government accountable.

To distinguish this story from this dissertation's argument, I am not claiming that Hu's or other PBSC members' beliefs had no influence in framing and contextualizing a logic of generating credibility among *Weibo* audiences, only that their worldview about the legitimacy of popular online participation in addressing societal problems not have substantial explanatory power for individual cases of non-censorship. Here, then, lies the difficulty with belief-driven explanations in explaining short-term tactical behavior: since belief change is thought to evolve over a long period of time, such explanations generally are of limited use in explaining fast-moving variation, such as decisions to censor, that must occur within minutes or hours, or a few days at most after some breaking event. PBSC members' personal views may well have mattered for affecting the general level of openness they were willing to tolerate, but such beliefs would offer little specific guidance as to what they (or their subordinates) should actually do case-by-case.

A second, similar 'non-rational' explanation has less to do with individual leader beliefs, than with widely shared attitudes within the Party toward certain issues and the legitimate scope about which discussion of each could be permitted. For example, the leadership collectively might be inclined to tolerate online discussion critical of the state on nationalist issues (e.g. related to opposing perceived Japanese provocations) than on domestic issues simply because elites sympathize

with netizens' anger on this issue in a way they do not regarding corruption, the environment, or other domestic areas. They then would set aside (or never take up) rational calculations, or at any rate these would be 'trumped' by more affective concerns. If elites were clear enough on what 'sympathetic' exceptions they were willing to allow, then, they could give appropriate standing orders to subordinates, allowing for quick reactions to breaking events.

While such an argument cannot be entirely refuted, two points are worth making. First, this explanation is again not inconsistent with a rational account. Here, I adopt Alastair Iain Johnston's (1995) standard for ideational explanation, which is that analysts who attribute causality to cultural or ideational variables ought to work diligently cross-temporally and cross-spatially to generate testable predictions that differ from those in rational models, since empirically the two may be (and often are) indistinguishable.[20] Elites may well have been sympathetic to angry online citizens, for example, as a blanket practice during any and all flare-ups with Japan. This does not mean that they then abandoned all cost-benefit analysis rooted in potential consequences for the Party's survival. In practice, distinguishing between the affective dimension of elite responses, and attributing a rational logic to their actions is likely to be intractable short of in-depth interviews with Politburo-level officials, a feat almost no scholars in the field have achieved. The theory here does not attempt to argue that leaders' actual thinking followed certain proscribed lines, only that their behavior was consistent with the instrumental logic I elaborate. Insofar as leaders' beliefs may have shaped their behavior in ways that better explain micro-variation in censorship decisions, my argument is open to challenge.

[20]This, of course, is true for any good testing of alternative explanations in social science, but here I raise a common critique of constructivist accounts – their supposed unfalsifiability. See Johnston (1995).

Second and finally, in its starkest form (that mostly non-rational motives drive censorship decisions), the argument is inconsistent with what we know about Chinese media censorship generally, namely its speed, sophistication, and tactical complexity. King, Pan and Roberts (2013) observe a state able to censor Internet content with "large-scale military-like precision". Other authors (Brady, 2008; Stockmann, 2013) observe swift and precise message synchronization across a range of media outlets, especially newspapers and TV news. While as just noted, plausible non-rational explanations for such rapid tactical flexibility exist, an account firmly rooted in the command-and-control logic of survival-oriented bureaucratic rationality arguably better explains how the system works, as the systematic language of survival- or security-driven thinking fits nicely with the speed and rigor of censorship implementation that we in fact observe.

CHAPTER 2

A THEORETICAL FRAMEWORK FOR EXPLAINING SELECTIVE SOCIAL MEDIA CENSORSHIP

2.1 Introduction

Control over the media has long been an essential piece of any autocrat's toolkit. Yet the actual level of control rulers choose to exercise varies widely across regimes. Some regimes, such as Morocco, prefer co-opting elites to censoring potential opposition (Willis 2014). Others view rigorous press, TV and Internet censorship as paramount, with China reportedly employing two million Internet censors (Bennett and Naim 2015). Yet even within the most stringent of censoring regimes, puzzling instances of *non*-censorship of seemingly sensitive issues sometimes emerge. To give one example from the Chinese case, the 2015 documentary film *Under the Dome* (*qiong ding zhi xia*) about Beijing's notorious (and life-threatening) air pollution problem was allowed to be shown on video sharing sites and discussed on social media for almost a week before abruptly being censored, despite the fact that it placed responsibility on local officials for failing to regulate polluting industries. Another example was the 2014 investigation of former Politburo Standing Committee official Zhou Yongkang, the highest-ranking official ever to be accused of corruption. Prior to the July 29 announcement of Zhou's investigation, his name was blocked in *Weibo* search but immediately became searchable and discussable thereafter. *Weibo* immediately exploded with commentary about Zhou's misdeeds and President Xi Jinping's anti-corruption campaign generally, including more than a few comments that criticized the Party more broadly. Yet for a brief

period, at least some fraction of even these volatile comments went uncensored.[1] Both incidents occurred during the early years of Xi's tenure, a period widely viewed as exhibiting decreasing speech freedom for bloggers and civil society alike.

Why does the Chinese government censor many but not all such incidents? This question goes to the root of how CCP elites have attempted to seize the 'commanding heights' of social media, and to selectively tighten and loosen their grip on platforms like *Weibo* in an attempt to retain and strengthen their hold on power. While cases of entirely predictable censorship abound — self-immolations in Tibet, unrest in Xinjiang, the activities of political dissidents — non-censorship is far less intuitive. This chapter builds off the premises introduced in Chapter One to develop a theoretical framework for understanding Chinese state motivations in varying social media censorship across issues and time. The plan for chapter is as follows. The next section reviews formal and non-formal theoretic literature on media censorship under autocracy generally and Internet control specifically. Next, I introduce the framework's basic setup and assumptions. Third, I develop its main dependent variable of censorship and key independent variables. Fourth and last, I use the *Under the Dome* case to illustrate how the framework's variables interact to generate predictions regarding censorship, as a prelude to the case studies in Chapters 4-6.

[1]Source: The Australian. July 31, 2014. "Web explodes as censors pull back."

2.2 Authoritarian Input Institutions and Information Control

Today's 21st Century autocratic rulers rely on far more than mere repression in order to stay in power. As noted early on by Geddes and Zaller (1989), authoritarian regimes do care about maintaining popular support. This is especially true for "resilient authoritarian" states (Dimitrov 2008) that have built institutions for maintaining popular support among various constituencies. A growing body of recent work has argued that authoritarian regimes employ quasi-democratic institutions strategically to bolster their position. Examples of such institutions include legislatures (Gandhi and Przeworski 2007; Gandhi 2008), courts (Liebman 2011), and media (Egorov, Guriev and Sonin 2009; Gehlbach and Sonin 2014; Lorentzen 2014). Work that focuses specifically on China has also examined how the CCP utilizes such institutions in a one-party context (Shirk 2011; Zhao 1998; Stockmann 2013). Despite an emerging consensus that quasi-democratic institutions increase incumbent regime stability, however, theories differ as to the specific mechanism through which they do so. One argument (Gandhi and Przeworski, 2007; Gandhi, 2008) is that forums like legislatures co-opt potential opposition by providing would-be critics a voice and access to the resources of the system. Other work focuses not on representative institutions, but on other conditions usually associated with democracy such as the presence of a vibrant civil society. In "consultative authoritarianism" (Teets 2013), civil society organizations represent societal interests to the state, while the state tolerates and even encourages select organizations that address social needs (social services, environmental protection, etc.) that the government has not met, but who do not directly challenge state authority.

One key divide in classifying theories of authoritarian input institutions is whether they emphasize horizontal accountability (rulers' responsiveness to other elites) or vertical accountability (responsiveness to ordinary citizens). Of these two types, work on why autocrats sometimes permit more open media is relevant to the latter. Yet such accounts vary widely in the extent to which they view tolerance of more open media for narrowly instrumental purposes, or as part of a broader effort at gaining legitimacy. On the more instrumental side, scholars have focused on the so-called "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. And even the leader's own agents, such as lower-level officials, have incentives to distort or withhold reporting on popular grievances since these often reflect poorly on the subordinates' own performance. Leaders can circumvent these barriers by allowing some media freedom, but this also allows potential opposition to use these channels to mobilize.

A growing body of formal-theoretic work (Egorov, Guriev, and Sonin 2009; Whitten-Woodring and James 2012; Gehlbach and Sonin 2014) has considered these trade-offs in-depth. Various models consider relative media freedom as a means to monitor local officials (Egorov, Guriev and Sonin 2009; Lorentzen 2014), as a key determinant in how citizens perceive "bad" political news (Shadmehr and Bernhardt 2015), and free media's effect on enabling citizens to publicly express preferences and the consequences of this for regime stability (Chen and Xu 2016). These models have been useful in establishing the key point that autocrats face costs as well as benefits in deciding to more tightly control information

flows. They also have disaggregated the multiple mechanisms through which the censorship level affects public opposition – directly by controlling citizens' ability to coordinate and mobilize, and indirectly by affecting their perceptions of regime strength, competence, or the degree of fellow citizens' dissatisfaction.

Of relevance to China, these models resonate with a vast literature on how the CCP stresses control over news media and the Internet as a crucial factor in "stability maintenance." Initial work on the Chinese media in the 1990s and 2000s found that even as market forces incentivized the Party to commercialize news outlets and to relax direct editorial control (Lynch 1999; Zhao 2000; Stockmann 2013), leaders used the propaganda system to make sure that publications generally served to reinforce (or at least, not undermine) the Party's authority. In the 2000s, the rise of Internet use in China injected a new dynamic into the mix. Broad analyses by Zheng (2007), Yang (2009), Morozov (2011), MacKinnon (2012) and others split on whether online spaces like blogs would enable social forces to challenge the state or enhance the latter's control in China and elsewhere, but all agreed that at least the potential existed for the state to co-opt and become dominant in the digital sphere. While some subsequent empirical work (Shirk 2011; Esarey and Xiao 2011) has indeed found that the Internet has empowered society to challenge specific state policies (though not to mount broad-based opposition), on balance these contributions have emphasized the state's ever-more robust control. Several studies (King, Pan and Roberts 2013, 2014; Bamman, O'Connor and Smith 2012; Zhu *et al* 2013; Fu, Chan and Chau 2013; Ng 2014) have shown Internet company censors' ability, under state supervision, to quickly and thoroughly remove (or make un-searchable) undesirable content from the Web, particularly from the high profile social networking services that increasingly predominate.

These findings, along with some theoretical accounts (Little 2016; Hassid and Sun 2015), particularly stress the CCP's specific interest in thwarting collective action organized online rather than indiscriminately suppressing all critical commentary. Yet while suppressing collective action is undoubtedly leaders' top priority, it is far from the Party's only interest in online space. A less instrumental and more legitimacy-based explanation for leaders' desire to shape the information environment is that they view media and especially the Internet as a tool to improve governance, that is, to both identify and address citizen demands and to better communicate (and possibly receive feedback) on specific policies. Some scholars believe that mechanisms such as local government Web pages for collecting citizen comments and other structured online forums are evidence of what He and Warren (2011) call "deliberative authoritarianism", a sort of digital public sphere where a limited degree of what Lewis (2013) terms "rational-critical deliberation" about policies can take place, at least for society's more educated and empowered individuals. While recognizing that the Party retains ultimate authority over the extent of online discussion, these works generally assess CCP efforts at promoting public deliberation as a genuine attempt to gain popular support for policies. Yet other work (Nathan 2003; Truex 2014) is more skeptical. Truex (2014) posits that in practice China has a very limited form of consultative authoritarianism that involves the state setting up constrained input mechanisms, such as public comment sessions and online feedback portals. The CCP may collect comments, but it often does not act on or even respond to them. And in an experimental study of local government responsiveness to citizen suggestions through both formal input channels and the Internet, Meng, Pan and Yang (2014) find that while local leaders are normally responsive to suggestions through both channels, responsiveness to online suggestions declines when officials perceive heightened social tension with

the local population, a result that points to the potential limits of authoritarian 'consultation'.

While both formal models about the trade-offs of freer media, and the authoritarian quasi-democratic institution literature have meaningfully contributed to understanding the Chinese state's approach to Internet censorship, neither has focused on another essential function leaders consider new media to possess: its utility as a channel for *persuading* the public of the Party's right to govern and the correctness of its policies – i.e. propaganda. However, in recent years a number of contributions have arisen to fill this void, building on prior analyses of propaganda's role in traditional media (Brady 2008; Stockmann 2010; 2013; Stockmann and Gallagher 2011; Zhang 2011; Lu 2014). Work focusing specifically on online propaganda has noted the Party's attempts in recent years to physically involve cadres and volunteers online in spreading pro-Party messages, from "cadres as bloggers" (Esarey 2015) to the so-called "Fifty Cent Party" (Han 2016).[2] Many journalists (Bandurski 2016; Lam 2013) have also noted the CCP's renewed emphasis since around 2011 on extending propaganda efforts to the online sphere. Most of these accounts view the CCP's purpose in its online push as a genuine attempt to make propaganda more persuasive to online audiences.[3]

Finally, one last factor when considering various aspects of information control and propaganda in China is the growth of Chinese microblogs. As discussed

[2]The "Fifty Cent Party" (*wu mao dang*) derives its name from the legions of ordinary citizens (often, college students and young people) allegedly paid *wu mao* or fifty Chinese cents for each pro-government post they write. It is pejoratively used to refer to any online poster that appears to be mindlessly spouting pro-government slogans, whether for financial or ideological motives.

[3]See, however, King, Pan and Roberts (2016). The authors show in their sample of social media posts attributed to "Fifty Centers" that the posts appear aimed at silencing critical discourse by flooding online space with pro-Party slogans and comments, rather than attempting to debate with and persuade the other side. However, this finding has unclear out-of-sample validity, and even if valid for "Fifty Cent" comments, does not disconfirm that the Party might have broader persuasive goals for its online strategy.

in Chapter 1, microblogs in China (as elsewhere) are characterized both by their technological characteristics – especially their ability to virally spread information and the ease with which anyone can post – and by their demographics, which skew educated, urban and wealthy, and include large contingents of journalists, celebrities and professional commentators. Numerous scholars have noted the potential for such unprecedentedly open-access yet elite-oriented platforms to challenge the state's news and commentary hegemony. Tong and Lei (2013) refer to this challenge as a "war of position" where the state's usual hegemony in media narratives is punctuated, in microblog space, by citizen narratives of "counter-hegemony" (p. 296) that call attention to injustices. Noesselt (2013) views the state as under so much pressure from microblogs that it now attempts to "base the political decision-making process on strategic calculations intended to be reflective of public online opinion" (p. 449), i.e. to respond to breaking incidents in ways intended to appease microblogger sentiments. And Xiao (2011) views *Weibo* as exemplifying the "cat and mouse game" between censors and Internet activists, in which activists sporadically gain the upper hand and expose abuses and scandals. Yet even though *Weibo* clearly has done much to expose the Party-state's shortcomings and to generate reform pressure, other authors (see Gunitsky 2015) have identified how China and other authoritarian regimes have quickly moved to co-opt social media to serve multiple needs such as information-gathering about potential opposition and mobilizing regime supporters.

2.3 Explaining Social Media *Non*-Censorship: Four Key Factors

The upshot of these diverse findings is that the Chinese Internet and particularly social media can be characterized as a "double-edged sword" (to use an oft-cited metaphor), but one which the state wields and where the edge pointed toward any social mobilization that might challenge state authority is sharper than the one facing the state itself, having been dulled by leaders' efforts to reduce the chance that allowing some degree of online freedom might backfire. In Chapter Three I discuss how such firm state control has been achieved, but for now assume that leaders have substantial ability to set both the terms and the extent of online political discussion. If this is the case, the next question might logically be: to what end? To what purpose do Chinese leaders apply such robust control? In the following sections, I go beyond existing theoretical and empirical findings on information control to elaborate a new framework that can be used to explain more episodic-specific variation in censorship than most work has taken up thus far, but that speaks to many of the above theoretical and empirical findings.[4]

A useful foundation for this effort is first to ask about the targets of such 'selective censorship': social media users. Chapter One established these individuals (and specific sub-groups such as celebrities and journalists) as highly educated even compared with other Internet users, and a key demographic who largely benefit from the existing order, but who are increasingly dissatisfied with pollution, cor-

[4]I term my explanation a 'framework' rather than 'theory' because it can be applied in a context-specific manner to explain (and potentially, to predict) individual episodes of social media censorship, but does not claim to comprehensively explain the state's online censorship behavior. It is, however, theory-like in that it links various explanations (like *responsiveness benefit*) to the dependent variable of censorship, and because implications of these linkages are empirically testable and falsifiable.

ruption, rising housing costs and other issues. This group relies to a great extent on "grapevine" or alternative information sources to gain political news (Zhu, Lu and Shi 2012), is more supportive of democratic norms and critical of the Party-State than the overall population (Lei 2011, p. 291; see also Tang 2006), and is more resistant to traditional propaganda (Geddes and Zaller 1989). Since these individuals are aware that the government, via its propaganda system, ultimately controls all publications and TV channels, they are likely to dismiss pronouncements via these media of its determination to reform as 'cheap talk'.[5]

Based on this group's characteristics, the challenge facing CCP leaders in deciding how to best manage social media to deal with their discontent is not well-explained by existing theories. An information deficit for central leaders about the sources of urban microblogger discontent along the lines of Lorentzen's (2014) model, for example, is unlikely because in contrast to issues affecting the rural population, such as land seizures by local governments, the sort of clean government and quality-of-life issues this demographic cares about are already widely understood both within this group and by government officials. In observing them on social media, CCP cadres may gain some sense of the acuteness of blogger demands, but they are not revealing problems of which they were previously unaware. Similarly, appealing to theories about various breaking online topics' collective action potential (King, Pan and Roberts 2013; 2014) cannot best explain the actions of this group because they are much less likely to protest compared with more disadvantaged groups, although as I note later on, collective action potential is still relevant in cases where they might be willing to physically mobilize. Rather, as introduced in Chapter One, I argue that leaders' task is to tighten and loosen

[5]Stockmann (2013) finds that market media are more credible to a range of audiences. My argument is not inconsistent with this finding, only that the highly-educated individuals who frequent *Weibo* are likely to be the most resistant to even market media, as these individuals have access to countervailing 'considerations' as per Zaller's (1992) model.

social media censorship in such a way as to *signal responsiveness* to this group's demands.

Here, I develop the intuition behind why allowing bloggers to speak out about some grievance during 'crisis' moments can be useful for elites to show responsiveness. Such a logic is ultimately rooted in the risky (from the regime's perspective) nature of shared public knowledge about regime shortcomings that often arises when citizens coalesce around some triggering event and are allowed to communicate freely. The risk that if left unaddressed, public speech that points a finger at the government for some widely understood problem might morph from merely demanding a policy response to more systemic opposition is precisely what credibly commits leaders to address the problem. The mechanism by which this occurs takes advantage of social media's unique capacity to facilitate common knowledge and information cascades (Granovetter, 1978; Kuran, 1991; Kuran, 1995, Lohmann, 1994; 2000). Take, for example, a hypothetical surge in online rumors saying that official X has been implicated in a bribery scandal. Assume that among the high-information *Weibo* public, a majority of individuals (say 80%) privately believe that official X is corrupt. However, they have not shared their views online, or if they have, censors up until now have quickly deleted such comments. Thus, even though four out of five netizens believe official X to be corrupt, most are unsure whether others share their belief.

Such a situation can persist in equilibrium for long periods of time, with large numbers of individuals with like preferences remaining ignorant of public opinion. The situation changes when some critical mass of online 'activists' decides to risk speaking out regardless of the prospect that they will be censored or even physically repressed; such a moment is often enabled due to some triggering event reported in

the news or spread via rumors.[6] Once the broader public sees activists speak, some of the more risk-tolerant among them will also take part, since they now know that at least some portion of the public shares their views. This, then, convinces even less risk-tolerant individuals that a broad majority is on their side, making them feel it is 'safe' to speak. If such speech takes place in a highly public forum such as *Weibo*, reciprocal knowledge arises – both the activists, and the new participants now know that each other knows that all speaking oppose official X. Critically, in this example top leaders who know official X is corrupt, intend to punish her, *and* need to show the public their intent also know that the public knows that they (top officials) are on the hook to follow through. The public assumes top leaders have such reciprocal knowledge, even without press conferences or official statements, precisely because people know that leaders made the decision to not censor information about official X in the first place. In a media environment where censorship of sensitive topics is the norm, censor inaction speaks volumes about where top leaders stand.

The reason why social media is an ideal space to facilitate such a cascade depends on the four groups of key actors and one key technology discussed in Chapter One. Taking each in order, social media companies like Sina benefit from the increased Web traffic (and thus, increase in user activity and potential avenues for monetization, like advertising) that breaking 'hot topics' provide. In fact, as a news-oriented product, *Weibo* in its first few years thrived in part due to the frequency and intensity of such speech bursts. Sina Corporation has an incentive to err on the side of risk in allowing a sensitive topic to propagate, until

[6] While some research (Roberts 2015) finds that netizen fear of state repression is not a primary mechanism through which censorship works, this finding was an experimental study in a low-profile and lower-stakes environment compared with *Weibo*, particularly *Weibo* blogging by public figures, who have in fact faced harsh repression since Xi Jinping's 2012 ascension.

explicitly ordered to shut it down.[7] Second are the Big V, who play the role of 'activists' in the above description. As discussed previously, the Big V, particularly entrepreneurs and businesspeople, are outspoken due to their unique position as private sector leaders and social role models, and due to their large numbers of followers. Third are journalists and editors, who enable Big V speech by reporting on breaking incidents or publishing unverified 'rumors' in *Weibo* space. Fourth are *Weibo* users, discontented with governance and quality of life and skeptical of government propaganda, whose response to Big V speech, placing the onus on top leaders to do something, ties leaders' hands and increases their credibility with this same *Weibo* public. Last is *Weibo* itself, which provides a 'public square' focal point (Patel, 2013) toward which all actors tacitly cooperate in directing their attention during breaking events.

This theoretical premise is related in some respects to at least two recent formal models. In an elaborate model the authors call an "information theory of dictatorship", Guriev and Treisman (2015) find that "censorship and co-optation of the elite are substitutes, but both are complements of propaganda." That is, spending more on propaganda can convince the general public of the regime's competence. However, for members of a (relative) elite such as microbloggers who are propaganda-resistant, the regime's only choices are censorship, or co-optation.[8] During breaking incidents which can expose government incompetence or corruption and for reasons I elaborate later on, censoring information such that even microbloggers do not draw a negative inference about regime competence is difficult

[7]There are exceptions, of course. Some topics are *a priori* banned via standing orders from regulatory agencies, usually in the form of banned keywords (common examples include discussions of Tibetan independence, and the 1989 Tiananmen incident). Others are not explicitly banned, but are considered so sensitive that Internet companies censor them to avoid crossing invisible lines – an example might be salacious gossip related to the personal lives of top-level (Politburo or higher) officials.

[8]Or violent repression, but in the authors' model this is the regime's most costly and least desirable option.

and can be costly. This leaves co-optation as the main option to placate this group. However, since these individuals are already relatively well-off in Chinese society, I argue that they would prefer to be "paid", at least in part, by non-material goods such as cleaner air, safe food and clean, efficient bureaucracy. Since citizen evaluations of whether these goods exist is *publicly* shared (sociotropic), rather than based on personal experiences with access to jobs or resources, the regime's task is somehow to publicly commit to microbloggers that it is making progress toward these goals. In a related model, Chen and Xu (2016) find that autocrats often prefer public deliberation to private polling because allowing public discussion "serves as a commitment device, ensuring that the government fully responds to problems that spur popular anger" (p. 5). The risk of this strategy is limited because citizens often disagree with each other when allowed to speak publicly, which undermines any collective action they might otherwise undertake. In cases where they do not, however, the government is forced to change policy (or carry out costly repression) because citizens have solved their coordination problem of becoming aware of each others' mutual discontent, thus greatly facilitating potential collective action.

2.3.1 Non-Censorship As Signaling: *Responsiveness Benefit* vs. *Image Harm*

Chen and Xu's model suggests that rational autocrats should be aware that not all collective speech, once allowed, will follow the same course. If citizens are sharply divided, then the government is under no pressure to change policy and also faces no unified collective threat. My argument differs from theirs and other theories, however, in that I consider various possibilities for what precise

common inference a unified online citizenry will draw, and the relative danger to the state of different inferences. Specifically, I consider whether microbloggers will become (relatively) unified around the sentiment that the government *must fix* the problem (but is not systemically incapable or corrupt) versus the view that the specific problem represents the "tip of the iceberg" of a government unable or unwilling to address citizen needs. I term the government's expectation that the former will prevail its expected *responsiveness benefit* and the latter its expectation of *image harm*, i.e. its public image as unified and capable of responding to citizen demands and realizing policy changes.

This focus on what *specific* shared knowledge citizens can be expected to generate resonates with many of the formal and non-formal theories above that focus on leaders' perceived competence, such as their use of consultative forums for citizen feedback to appear as effective problem-solvers. It differs from these theories in its explicit emphasis on *non*-censorship as the knowledge-generating mechanism. Without relying too much here on game-theoretic language, perceived competence can reach two different equilibria. In one equilibrium, citizens believe the government to be basically competent and effective, and that they have locked leaders into addressing the problem; in this case the government reaps the benefit of being seen as problem-solver. In the other, citizens view openness not as a credible gesture of government intent but rather as weakness – that the problem is so out-of-control that it is impossible to cover up. Of course, in reality the "common knowledge" citizens reach is rarely ever purely the first or second equilibrium. What leaders do regarding censorship, then, depends on their expectation about the overall tendency toward one of the other.[9] I thus define a composite factor, leaders' expected

[9]The distinction between the two is not primarily about 'negative' versus 'positive' speech, or criticism of the Party versus lockstep support. Rather, it hinges on leaders' judgment as to whether particular *framings* of an issue or incident, even if they give rise to sharp anti-Party criticism, will allow the Party the opportunity to respond constructively. In other words, there

credibility payoff, as the difference between *responsiveness benefit* and *image harm*:

$$credibility\ payoff = responsiveness\ benefit - image\ harm \tag{2.1}$$

Before proceeding further, I should clarify a few aspects of the government's expected behavior. First, I am not claiming that leaders either have perfect foresight with respect to calculating *credibility payoff*, or even that they are particularly good at anticipating the direction of online reactions to breaking events. Neither do the officials below the central leadership tasked with making day-to-day censorship decisions likely contemplate their decision explicitly in the utilitarian terms laid out here, relying instead on analogues to similar incidents, and hunches drawn from their experience working in China's media and propaganda system. However, once an incident breaks on social media and draws senior officials' attention, they know they have to quickly size up the situation and either order increased censorship, or deliberately let discussion proceed.

Second, just because a theoretical "balance" between *responsiveness benefit* and *image harm* exists in Equation 1 above does not imply that the balance will often, or even normally be positive in the real world. Indeed, as discussed in Chapter 1's scope conditions, censoring officials will only perceive the former as outweighing the latter on issues and during crises where they think the *Weibo* public absolutely demands that they acknowledge the problem and take action; the balance can thus only be expected to be positive in egregious revelations concerning a handful of major issues that are of vital interest to this group *and* that leaders view as not violating the Party's political bottom line. While the actual sources of *Weibo* user discontent do matter for which issues generate *responsiveness benefit*,

must exist a 'way out' for leaders.

China's leaders still play a decisive role: only those issues where leaders think they must appear accountable to the *Weibo* public qualify. *Credibility payoff* is, therefore, *partially* endogenous to leaders' own perceived incentives. On the one hand, leaders' thinking on which issues they need to appear accountable cannot be totally divorced from the *Weibo* public's preferences – if it were, the claim that leaders have a rational need to demonstrate responsiveness would not hold. On the other hand, the recognition that reform (or better policy implementation) in a given area is necessary must begin with the CCP leadership itself: Party elites must first come to believe that their legitimacy truly rests on convincing the public they are serious about addressing a problem.[10]

2.3.2 *Collective Action Risk*

Although this chapter's novel contribution is proposing a calculus where dynamic changes in social media censorship depend on leaders' assessment of *responsiveness benefit* versus *image harm*, the "theory of collective action potential" promoted by King, Pan and Roberts (2013; 2014) is an additional factor that deserves consideration.[11] In two papers, the authors convincingly show that the Chinese state censors online comments that "represent, reinforce, or spur social mobilization, regardless of content" (2013, p.1). This theory differs substantially from the explanation reflected in *responsiveness benefit* and *image harm*, in that it says that the specific common knowledge or frame generated by bloggers – whether bloggers

[10]While in practice there is substantial overlap between educated, urban citizen concerns and issues the leadership is willing to consider, one could imagine some issues of concern to these individuals – perhaps a movement for stronger legal protections of free expression, for example – where the Party's political imperatives would conflict with and could 'veto' citizen desires.

[11]Throughout the dissertation I refer to 'collective action risk' instead of 'collective action potential' to emphasize that what a forward-looking state really cares about is the *expected* risk that an incident will lead to collective action.

simply seek remediation of the problem or are engaging in a broader anti-system critique – is not what matters. Rather, the regime censors any posts that suggest collective action *on the ground*, even if these are limited to the specific problem. Numerous related articles analyzing what keywords are most often censored (Zhu *et al* 2013; Ng 2014; Fu, Chan and Chau 2013; Bamman, O'Connor and Smith 2011) similarly find that the regime censors topics related to collective action. However, these studies also uncover a range of terms blocked on *Weibo* and elsewhere online that seem unrelated to any form of real-world mobilization. Oft-cited examples include the names of top leaders, broad criticism of the Party-state, and natural disasters that do not have any obvious potential to trigger unrest. Similarly, a recent analysis of leaked censorship directives issued by the Party's propaganda bureaucracy (Tai 2014) finds that censors "ban news that directly threatens the legitimacy of the regime" (185), regardless of collective action potential.

In weighing the utility of the theory of collective action potential, I propose that we need to more carefully consider its appropriate scope. First, a methodological point: King, Pan and Roberts' landmark (2013) study excluded *Weibo*, sampling from blogs, BBS, and smaller websites. While there is no *a priori* reason to believe such a sweeping theory would not apply to *Weibo*, this omission leaves open the possibility that the logic of censorship may differ across platforms – in fact, the broad range of censored keywords and topics unearthed by other studies raises the possibility that on *Weibo* a more nuanced logic may be at work. That said, there do exist numerous examples on *Weibo* of collective action-relevant topics being censored.[12] Clearly, given what we know about the Chinese state's longstanding

[12]Examples from 2012 included localized protests over proposed chemical plants (a molybdenum-copper plant in Shifang, Sichuan), dissident activities (blind activist Chen Guangcheng's flight to the US Embassy compound in Beijing), and natural disasters with collective action potential (a July rainstorm in Beijing that led to heavy flooding), to name just a few.

suppression of protest-related information, the theory of collective action potential is at least partially applicable to social media. I therefore adopt King, Pan and Roberts' version of this theory as one variable in the framework, with two caveats. First, I argue that their account does not explain all, or even most variation in what is censored on *Weibo*; in particular, it cannot explain cases where a trending topic's link to collective action is tenuous or nonexistent, but is censored.

Second, and conversely, there exist at least some cases that showed high collective action potential but were not censored, at least for some period of time (Cairns and Carlson, 2016). Thus, an incident's level of collective action potential does not invariably determine the outcome. To be sure, as the risk of widespread collective action to the state surges in rare but dramatic instances of social mobilization, such as anti-foreign protests, the state is eventually likely to censor social media. Above a certain level of volatility, leaders simply will not tolerate online speech that might spur massive anti-regime opposition. Yet treating collective action potential as deterministically predictive of censorship is simplistic.

Before going further, it is important to clarify the distinction between *image harm* and *collective action risk* (or 'potential'). Both appear to involve threats to the state, and are related in the sense that prolonged or widespread collective action could demonstrate regime incapacity, while sufficient *image harm* could eventually invite collective action. However, the two differ in that collective action potential concerns the possibility that if left uncensored, bloggers will directly and immediately coordinate real-world actions such as protests. It refers to Little's (2016) "coordination" effect of social media rather than merely generating common knowledge about the regime.[13] Collective action poses an immediate and grave risk

[13]Of course, shared knowledge of a problem is necessary to motivate people to take part in collective action, but not sufficient to make it happen.

from the state's perspective, and is only tolerable in highly circumscribed forms, such as some nationalist protests. *Image harm*, on the other hand, is survivable in the short term and worthwhile if leaders think that online sentiment trends appealing to the state as problem-solver will outweigh minority currents that question regime legitimacy.

2.3.3 *Visible Censorship Cost*

The first three variables are each defined as either the benefit leaders expect to receive or the cost they expect to pay if they do *not* censor. Since estimating all three involves the difficult exercise of predicting what the *Weibo* public will do, if *responsiveness benefit* does not clearly outweigh the other two factors, then why would Beijing ever do anything other than censor? One might think the censorship option would be costless, i.e. that Beijing's payoff would simply be zero. However, in an experimental study, Margaret Roberts (2015) found a meaningful distinction between "visible" and "invisible" censorship. "Invisible" censorship prevents harm to the state's image via what Roberts calls "friction" — increased physical difficulty in locating information about an incident, even if a user already suspects something is happening. If netizens do not know exactly what they are being prevented from seeing, the state bears no extra cost. In contrast, visible censorship has been hypothesized to work via fear, warning netizens to desist from creating or seeking sensitive information or suffer unspecified consequences. Roberts finds, however, that in reality netizens are not intimidated by observable censorship; rather, it increases their desire to acquire and spread forbidden information and even to generate more sensitive comments themselves.[14] Based on Roberts' research, I

[14]Applying Roberts' finding to the framework does not contradict Footnote #6 because her

thus define *visible censorship cost* as the government's expectation regarding the likelihood that censorship will be visible and trigger this negative effect, which can both make subsequent efforts at censorship more difficult, and negatively influence citizen perceptions of government capacity and honesty, since bloggers think leaders are trying to cover up a severe problem.

One key factor in whether the state will incur this cost is if bloggers have prior knowledge of the issue, making them more on the lookout for repeat instances. For example, at any given moment many *Weibo* users have prior knowledge and views regarding China's maritime disputes with Japan, making breaking information about flare-ups very difficult to suppress. On the other hand, bloggers might have less reliable (or simply less) information about unexpected disease outbreaks (such as SARS' 2003 spread throughout China), which would facilitate attempts to censor any information that did surface before bloggers really came to understand the problem, or to dismiss as "rumors" any information that did leak out. A second factor affecting censorship visibility is the state's technical ability to control the information environment. One implication of Roberts' work is that the state, ideally, would like to make censorship entirely invisible – so subtle and yet comprehensive that netizens are unaware it is even occurring. In fact, we do see Sina *Weibo* using ever-subtler means to 'hide' ongoing censorship (see Ng, 2014). Whereas in *Weibo's* first few years, keyword searches for a censored term would yield a telling message, "according to the relevant laws and regulations, results for the requested term cannot be displayed", in 2014 *Weibo* changed the 'error' message users would receive for a blocked term – for many sensitive terms, *Weibo* now returns 'no results found'.[15] However, even with technological refinements

experimental results, in my view, are more likely valid for *Weibo* bloggers as a whole, than for the Big V 'activists' referenced previously.

[15]Use of "according to the relevant laws..." continues, however. See Ng (2014).

like these, the state still does not have total control. News about some incidents may spread so virally that some users are able to view the message before post deletions and keyword blocks are put in place. And on some issues, Hong Kong, Taiwan, or international media may bring attention to a particular incident, causing it to reverberate within China (including in some mainland news outlets). The key implication of including censorship visibility in the framework is that as long as censorship is at least partially visible, it always carries a cost. That is, even if the other three factors are zero or cancel out, Beijing rarely receives a net utility of zero if it censors, though it may pay an even larger negative cost if it does not.

The rest of this chapter shows how the above four factors interact to produce specific predictions regarding censorship. As Chapter 1 discusses, the factors (and censorship) are dynamic and variable both over time, and within various sentiment categories that rise and fall during a given online episode. Before doing so, though, I first more rigorously define the dependent variable of social media censorship, which proxies for the broader concept of freer or more restrictive information and media control.

2.3.4 The Dependent Variable: Social Media Censorship As Measure of Overall Information Control

Throughout the dissertation, social media censorship proxies for a broader concept: authoritarian information control. There are good reasons for this, namely that censorship – the redaction of publishable content by governmental or other official authority – is of course a primary means of information control in China

and elsewhere.[16] Yet states use many other means of shaping information flows, including harassing or arresting writers and bloggers, financially incentivizing publications to generate only pro-regime content through investment or ownership, and website blocking. All of the above are present in China and have received substantial attention. However, on *Weibo*, censorship in the form of deleting user posts is a frequently used means of restricting information.

I consider the level of censorship – specifically the percentage of *Weibo* posts deleted for a given topic and day – to gauge tight or loose information control. To measure this level I 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, and 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 my 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). Using this data, I define the censorship rate in Chapters 4-6 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 is not perfect, it is the best available method to get some measure of the speed and volume of censorship. However,

[16]There exists no exact translation of 'censorship' in Chinese. The two closest terms in Chinese usually used are *shencha* and *jiancha* which literally mean 'to examine and check/look into', and 'to check up on and check/look into'. Both have the connotations of examining, checking, or inspecting material. While many similar terms, such as *diaocha* (to investigate) and *shenpi* (to audit) do connote the exercise of authority, they all (including *shencha*) have the sense of a natural and necessary function, as Westerners might view auditing a company's financials or the police investigating a crime. Nowhere present is the sense of abnormal, heavy-handed control or an infringement of free speech. To complicate matters further, much state discourse that refers to 'censorship' actually does not use *shencha*, preferring terms like 'Internet management' (*wangluo guanli*).

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 (Cairns and Carlson 2016), I 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, with further 'face validity' tests of the measure in Appendix B.4.

Relying on a measure of *Weibo* post deletions as indicator of state bureaucrats' (and ultimately, senior Internet-bureaucracy officials') short-term information control intentions appears unrealistic at first glance given the fragmentation and size of the Chinese state, not to mention the fact that Internet companies actually manage post deletions in-house rather than state agents doing so directly. In Chapter Three, however, I show that recent reforms to China's Internet censorship system, particularly as they affect social media, have made inferring state intentions from *Weibo* censorship behavior more plausible than most analysts of the Chinese bureaucracy would typically grant. First, though, this chapter's final section ties together *credibility payoff* (*responsiveness benefit* and *image harm*), *collective action risk* and *visible censorship cost* into a framework for predicting censorship outcomes across issues, sentiment categories, and time.

2.4 Putting the Factors Together: Explaining (Non-) Censorship In Specific Cases

Chapter One introduced the unit of analysis for this study as "breaking inci-

dents in issue areas resonant with *Weibo* users." Each incident, in turn, is nestled within a particular issue area which can be assigned scores on *credibility payoff*, *collective action risk* and *visible censorship cost*.[17] This issue-level variation is what led me to designate Chapters 4-6 as "easy", "hard", and "medium/mixed" cases for selective non-censorship: air pollution as an "easy" case of non-censorship (Chapter 4), the investigation of Politburo official Bo Xilai as "hard" (Chapter 5), and the 2012 Diaoyu/Senkaku islands protests as "mixed" (Chapter 6). Yet I also expect censorship to vary in theoretically meaningful ways *within* each incident/issue, and to do so across two dimensions: over different sentiment categories, and over time. Across-category variation in censorship and the associated values for the framework's key variables can be quite nuanced depending on the specific category characteristics, and in Chapters 4-6 I consider how the characteristics of individual categories – for example netizens' belief that the central government is not sufficiently "tough" on defending China's territory against foreign incursions – led to my specific codings of the three variables. In general, however, sentiment categories or trends can be divided into two groups: those that while critical, offer leaders the chance to pose as problem-solver, and those that portray the state as incapable or even compromised and corrupt, i.e. irredeemable.[18]

Second, both censorship and the key independent variables can be expected to vary dynamically over time. *Credibility payoff* should be greater during perceived periods of relative stability and lower during perceived instability. Both events planned in advance by the Party that are known for being sensitive, such as leadership transitions and major Party conferences, and unanticipated events

[17] To simply analysis, throughout the dissertation's empirical sections I refer to *credibility payoff* rather than *responsiveness benefit* or *image harm* individually.

[18] Any issue analysis will also include "neutral" categories that do not neatly fall into A or B, such as mere reposts of online news. This is not an issue for conceptualizing the cases so long as such neutral sentiments do not predominate.

like popular uprisings in the Middle East can be expected to influence leaders' estimation. These events also may or may not increase perceived *collective action risk* – for example, the 2014 Occupy Hong Kong protests would be expected to do so through fears of "copycat" protests on the mainland, but "Arab Spring"-like activity elsewhere might not, at least not acutely. Finally, periods of increased domestic and international media reporting during major events might increase *visible censorship cost*, although this increase would likely be outweighed by other factors in favor of stronger censorship.

At this point, one might ask how the analyst is supposed to arrive at *ex ante* categorizations for each episode/issue, sentiment category and time point across all factors, and generate a prediction concerning censorship. Certainly, deep familiarity both with Chinese politics and with the specific issue area under consideration is required to attempt such a categorization. That said, the exercise is not nearly as *ad hoc* as it might appear, as in practice there is fairly widespread consensus in the field on how Chinese leaders are likely to 'score' certain variables: for instance, issues that are salient to urban, middle-class citizens such as air pollution and nationalism are likely to be much more visible if censored than lower-level corruption or ethnic unrest. Similarly, what the leadership views as generating high *collective action risk* is to some extent known by outside observers, with China-Japan confrontations and so-called 'NIMBY' ("not in my backyard") protest episodes high on the list.

Formulated concretely, then, the categorization exercise (though not wholly without controversy) is feasible given detailed knowledge of Chinese politics. Yet it is still admittedly a demanding exercise that requires the analyst to holistically analyze each case and then make a prediction, based on direct qualitative and

quantitative social media data analysis, official interviews, publicly available information, and/or the analyst's own judgment how to score each of the above four variables by issue, category and time, before proceeding to measure the censorship rate and assess the framework's 'fit' for that case. To illustrate how such a scoring exercise might unfold and to prepare the way for Chapters 4-6, the remainder of this section does so with respect to the example of the *Under the Dome* documentary.

2.4.1 Predicting Censorship: The Case of *Under the Dome*

The 2015 documentary *Under the Dome*, an exposé about the sources of China's air pollution problem and its health consequences by well-known former China Central Television host Chai Jing, went viral on Chinese video sharing sites immediately after its February 28 release and brought unprecedented attention to both the dangers of air pollution, and a lack of action on part of those responsible. The film received over 300 million online views in just a few days, making it one of the most widely viewed videos in Chinese Internet history. Its release occurred just a week prior to the start of China's so-called "Two Meetings" (*liang hui*), which are the annual meetings of the National People's Congress (NPC) and Chinese People's Political Consultative Congress (CPPCC), two of the country's most important (and nominally representative) legislative and consultative organs. The NPC and CPPCC's annual meetings include delegates that represent a range of business and social interests, and like similar bodies in other authoritarian contexts, serve as both a feedback mechanism for policies introduced by the central Party, and a means for integrating and co-opting elites outside the central Party structure. While I have no definitive proof that Chai intended the film to influence

the NPC/CPPCC agenda, the timing of its introduction during what normally would be considered a highly sensitive (and censored) period lends support to this interpretation. The film itself contained several provocative themes, including the human toll of air pollution (Chai discusses how her own unborn daughter developed cancer *in utero*, and blames air pollution), and attacks on those Chai believes to be responsible including China's state-owned petroleum refining monopoly, Sinopec, whom she accuses of failing to meet international standards for clean gasoline.

Despite these sensitive points, however, the film was allowed to remain online for a full week after its release (and even overlapping with the Two Meetings' opening day), *Weibo* commentary appears to have been relatively uncensored, and even state-owned media were encouraged to publicize the documentary, with *People's Daily* posting the video and offering Chai an interview about why and how she decided to make the film.[19] This led international media outlets to argue that China's Ministry of Environmental Protection as well as media regulators had pre-approved the film.[20] Nonetheless, censorship still came beginning on March 2 with orders for media outlets not to report on the film, followed by the documentary itself being removed from video sites on March 7 and tighter censorship of social media commentary. This curious pattern of deliberate openness followed by sudden censorship begs the question: why allow the film to be shown (and discussed publicly) in the first place? While this dissertation does not undertake a full case study to generate and test hypotheses about this question, Tables 2.1 and 2.2 below give a rough example of how one might apply *credibility payoff* and the other variables to work toward an answer. Table 2.1 shows my estimate of how authorities would have scored each variable in the period prior to and during the

[19]Source: The New York Times. March 2, 2015. "Documentary on pollution stirs Chinese."

[20]I confirm this with interviews from multiple sources including a leading environmental NGO activist with government ties, and a well-placed source close to the film crew. See Chapter Three.

first day or so of the Two Meetings, before the Meetings' agenda was in full swing, while Table 2.2 shows estimates from March 7 forward. The choices of categories are meant to be illustrative, not exhaustive.[21]

Table 2.1: *Under the Dome* Predicted Censorship by Category (2/28-3/6, before "Two Meetings")

Category	Cred. Payoff	Coll. Actn. Risk	Vis. Cens. Cost	Pred. Cens.
Blame petroleum interests	Positive	Low	High	Low
Blame central government	Negative	Low	High	Medium

Table 2.2: *Under the Dome* Predicted Censorship by Category (after 3/6, during "Two Meetings")

Category	Cred. Payoff	Coll. Actn. Risk	Vis. Cens. Cost	Pred. Cens.
Blame petroleum interests	Negative	Medium	Very High	High
Blame central government	Very Negative	Medium	Very High	Very High

The two tables each contain the same two potential sentiment categories.[22]

[21]I have not examined (and for this project, have no intention to examine) any *Weibo* or other social media data in-depth about the *Under the Dome* case aside from cursory glances (while events were unfolding) at the main *Weibo* feed and people I follow. I only know about estimates of censorship from media reporting, which was not precise about exactly how much censorship occurred and what was censored. Thus, the independent variable codings in Tables 2.1 and 2.2 are based on my own prior knowledge of Chinese censorship in other cases, knowledge of the environment and issue area, and my own theoretical priors, and are not in danger of being 'fit' to knowledge of measured censorship outcomes from this case.

[22]The categories are examples of how an analyst might choose to score online discussion after studying the incident and reading *Weibo* text, and do not necessarily represent what I believe to be the definitive categories for this incident, a list that could only be obtained after in-depth

The first would be netizens who primarily viewed Chai's documentary as putting blame on Sinopec for being unwilling to refine cleaner gasoline, while the second would more broadly fault central leaders for not cracking down on contributors to pollution – they would view the central government, not Sinopec as at fault for failing to enforce regulations. In the 2/28-3/6 period, I coded the first as "Positive" *credibility payoff* because central leaders would likely have no issue with scapegoating what the Chinese public already viewed as entrenched and corrupt energy interests, and might relish the opportunity to be viewed as problem-solver in taking a hard line against these actors. However, they would of course less favorably view the second category of blaming central leaders themselves. I coded *collective action risk* in this period as "Low" because to the best of my knowledge, air pollution has not generated any on-the-ground protest activity. The problem's sources are diffuse and citizens lack any easy target for collective action.

Finally, I coded *visible censorship cost* as "High" for both categories.[23] Public awareness was undoubtedly much higher for *Under the Dome* than a typical online "breaking incident", which might attract millions but not hundreds of millions of views on social media. This would have been true for both issue categories, since for such a high-profile event once discussion began it would be difficult to filter out and censor only those comments blaming petroleum interests. The implication of this coding is that if the government chose to relax censorship, it would do so in spite of knowing it would inevitably have to incur this cost if it later reasserted control.

analysis. Typically, a coding exercise will, at a minimum, also require designating a "neutral" category for topic-relevant content without strong affective sentiments, such as mere reposts of news articles.

[23]Although theoretical cases exist where *visible censorship cost* could vary between different categories as a result of the public's much greater awareness of some categories versus others (for example, framings based on reported information versus unconfirmed rumors), often the same event will underlie all categories, making their *visible censorship cost* equal.

Turning to Table 2.2, I coded *credibility payoff* as "Negative" for the "blame petroleum interests" category and "Very Negative" for "blame central government". As time proceeded and the film accumulated more and more views, the theme that the central government bore some share of the blame became more pervasive. Additionally, the government had already publicly acknowledged the problem, with Premier Li Keqiang stating on the Two Meetings' first day (March 5) that "environmental pollution is a blight on people's quality of life... we must fight it with all our might."[24] After this acknowledgement, leaders likely felt they had done enough and did not stand to gain further from allowing public pressure to continue. Even talk that continued to focus on the petroleum interests was risky as it would become more easily linked to perceived central government failures to combat corruption and entrenched interests. Second, I coded collective action risk as "Medium". This would be an unusual coding for the air pollution issue area, where on-the-ground collective action had not occurred previously, but I considered it appropriate given the unprecedented public attention and anger generated by the documentary, which might eventually coalesce into citizen protests. Finally, I coded *visible censorship cost* as "Very High". Many netizens certainly took note that the government abruptly shut down discussion after allowing commentary and video shares great latitude for days, and this probably raised questions in netizen eyes about leaders' intentions. However, from the Party leadership's perspective, the increasing weight of the first two variables would have trumped this consideration.

As this exercise shows, the framework's factors can theoretically vary across each sentiment category and moment in time, but normally do not all do so at

[24]Source: The Guardian. March 5, 2015. "It is more than a documentary; as viewing figures rocket past 300m, officials seem to be taking a tolerant view of Chai Jing's film, which examines the issue of deteriorating air quality."

once. Instead, they tend to follow semi-regular trends that simplify the analyst's task of assigning a score to each category-moment combination. For example, for any given issue area and incident, *visible censorship cost* tends to increase over time, provided authorities initially allow discussion. And *collective action risk* usually applies to *all* categories within a topic (see King, Pan and Roberts 2013, p. 7) since censors usually have difficulty in distinguishing online sentiment trends related to collective action from those that are not; an exception would be if some sub-set of posts explicitly called for protests in the context of a much broader issue discussion, as happened during the 2012 Diaoyu/Senkaku islands protests. These patterns mean that the critical difference between why some categories are censored will often depend on their *credibility payoff*, allowing analysts to first assign scores that are consistent within time periods for the other two variables, and then to focus on assigning *credibility payoff* to each individual category.

Chapters 4-6 each delineate relevant categories within their respective incidents, score them in a manner similar to (but in greater depth than) the above example, and make category-moment specific predictions of censorship before delving into the WeiboScope data to develop measures for each. While no single issue or category can conclusively support the validity of *responsiveness benefit* and an associated selective censorship logic, patterns of variation both across issue-incidents and within them are consistent with the theory. First, however, it is necessary to lend empirical support for the claim – until now an assumption – that the Chinese state exercises sufficiently unified and top-down control over Internet companies to be able to meaningfully speak of "the state's" censorship strategy. Next, Chapter Three takes up this task.

CHAPTER 3

FRAGMENTED AUTHORITARIANISM? REFORMS TO CHINA'S INTERNET CENSORSHIP SYSTEM

3.1 Introduction

"The processes [in China] through which large-scale energy projects are decided reveal that the fragmented, segmented, and stratified structure of the state promotes a system of negotiation, bargaining, and the seeking of consensus among affected bureaucracies. The policy process in this sphere is disjointed, protracted, and incremental." – Kenneth Lieberthal and Michel Oksenberg, *Policy Making in China: Leaders, Structures, and Processes*, p. 3

Almost two decades after going to print, Lieberthal and Oksenberg's landmark study of policy making in China remains an important framework for understanding policy outcomes in the world's largest bureaucracy. Although the authors admitted the potentially limited scope of their findings due to selecting a sector (energy) more prone to bureaucratic fragmentation than many others, subsequent work (Lieberthal 1995; Li 1998; Mertha 2005; Brodsgaard 2006), to name just a few contributions, has since validated and expanded the bureaucratic model. More importantly for this dissertation, work on China's media and propaganda bureaucracy (Lynch 1999; Shambaugh 2007; Brady 2008; Stockmann 2013) has highlighted the differing interests of Party propaganda departments (at the central and provincial levels), provincial and central government agencies and media organizations. This model, along with the "rational" and "power/factional" poli-

tics accounts to which Lieberthal and Oksenberg compare their own argument, is today a fundamental building block of Chinese political analysis. Yet when they first introduced their model, the authors considered neither its applicability to the propaganda system, nor to media management generally — and certainly could not have foreseen the consequences the rise of the Internet, yet alone social media, would have for its appropriate scope.

The question then remains how applicable the bureaucratic model is not only to media policy and censorship, but in an Internet era characterized by the dominance of just a handful of companies located in major cities – an 'easy' regulatory target. To look ahead, some aspects of my findings do resonate closely with the bureaucratic model. For instance, Lieberthal and Oksenberg note that "there is usually a series of iterations in this [policy refinement] process, where the initial *zhengce* [policies] prove inadequate and are supplemented by ever more refined administrative orders..." (p. 26). Such an iterative, evolutionary process found strong support in my interviewees' descriptions of leaders' management of *Weibo* as an 'experiment' from 2009-11. Similarly, the authors note that "the emergence of a critical problem [e.g. a crisis or 'shock'] may capture the attention of the top leaders and force decisions to be made" (p. 30). This chapter's analysis of Internet *Xitong* (bureaucratic system) reform similarly stresses the vital exogenous importance of social media's rapid development in grabbing top leaders' attention and spurring policy reform.

Yet in many other respects, Lieberthal and Oksenberg's characterization of the "fragmented, segmented, and stratified structure of the state" simply does not work well for Internet regulation, and especially for attempts to regulate companies like Sina, Tencent, and Baidu that operate market-leading social media services.[1]

[1]The model may work much better for inter-bureaucratic and central-local divisions regarding

Of course, given the central importance of complex bureaucratic layering to implementation as well as policymaking in nearly all policy areas, caution is warranted in making such a claim — the burden of proof should be on researchers to show not only *that* such regulation is far more unitary and top-down than elsewhere, but *how* and *why* the normal conflicts to which other policies are captive do not apply to the Internet.

While this chapter cannot fully reconcile the bureaucratic politics model with Chapter One's necessary conditions for selective censorship – elite intentions to reform the Internet system and the bureaucratic logic of how they have done so – the findings below both advance theory and provide hard evidence in this direction. I identify three relevant factors in leaders' success: a) longstanding thinking among political elites as to the value of ICT control, as well as having made prior efforts; b) a symbiotic relationship between the state and Internet companies; and c) whether specialized agencies get the Internet management portfolio or whether other agents – especially media and propaganda agencies, and the state security apparatus – are given jurisdiction over online space. I argue that for China, the first two conditions have been necessary but insufficient because both existed prior to Xi taking power. China is fairly unique among states in the degree to which the ruling party has stressed centralized control over ICTs, going back to the telegraph under the Qing Dynasty (Zhou 2006) and before. And no other authoritarian state has as large and vibrant a domestic Internet sector as China's, and one so deeply invested in symbiosis with political authority.

But even if these conditions are present, which agencies hold the Internet portfolio matters for whether censorship policy will be adequately flexible and swift to

regulation of smaller websites and bulletin boards (BBS). However, all these have declined in importance not just on the Chinese Web, but globally in recent years, and are not this chapter's focus.

adapt to such a dynamic medium. A corollary to this third factor is that who gets put in charge of the Internet is inseparable from President Xi's broader struggle to consolidate power away from bureaucratic incumbents he views as opposed to his reform program. These three factors then suggest that the Internet sector may differ substantially from the policy fragmentation found elsewhere. This makes formal models' assumption of a unitary, rational state more plausible than many Chinese and bureaucratic politics scholars have been willing to allow.

The following sections undertake this task through a case study of leader attempts to re-centralize the Internet bureaucracy beginning around 2011. After briefly explaining data collection and research methods, I consider leader intentions to reform this system prior to Xi. Third, I address the symbiotic relationship between Internet companies and the state. Lastly, a comparison of pre- and post-reform bureaucratic structures reveals how concerted efforts to empower Internet "experts" at the expense of both China's existing propaganda system and the state security apparatus transformed China's censorship system from moderately strong, to very robust.

3.2 Data and Method

This chapter's primary data source is 57 targeted elite interviews I conducted in Beijing, Shanghai, Guangdong and Hong Kong in 2014-15 and summer 2016. The three mainland sites are home to nearly all of China's Internet media giants. Hong Kong was included as it is home to a number of journalists and communications scholars who study mainland censorship. Interviewees fell into one of three major categories: Internet company insiders, journalists, or media-oriented academics.

Due to the topic's sensitive nature and the restrictive political climate at the time of fieldwork, I did not record interviewees but relied on handwritten notes taken during and after each interview. To protect their identities, all interviewees are cited anonymously, with each citation giving only the interview number (by order conducted), city, and date, with limited background information given only where safe to do so.[2]

Interviews usually lasted about 1-2 hours and were as casual as possible to put participants at ease and invite them to share information on their own terms. Questions were semi-structured: I chose about 10 questions per interview from a loosely standardized list of several dozen, based on their anticipated relevance to the interviewee's expert knowledge, and to avoid excessive sensitivity that might provoke a non-cooperative response. My interviewee pool began with a few individuals reached via academic contacts in the U.S. and China, and grew through the snowball method; at the end of each interview I asked the participant to refer close friends or associates who might be willing to speak – typically, this led to 1-2 referrals of long-trusted contacts.[3] Through persistence, I was able to slowly build out the pool until I had reached over 40 individuals by the end of fieldwork.[4] Due to the political constraints prevalent in 2014-16 – an anti-corruption crackdown that heightened officials' fear and paranoia, as well as a specific effort beginning in late 2014 to 'rectify' (*zhenggai*) the behavior of Internet-relevant cadres, access to government officials was severely limited.[5] Nonetheless, I was able to speak

[2]Location codes: BJ = Beijing, HK = Hong Kong, SZ = Shenzhen, GZ = Guangzhou, SH = Shanghai.

[3]I attribute interviewees' typically limited number of referrals both to the topic sensitivity and the political pressure on media practitioners under Xi, and the topic's specific and technical nature, which may have led interviewees to carefully filter their contact lists for individuals they thought would actually be able to say something useful.

[4]I interviewed about ten exceptionally valuable participants more than once, giving a total of 57 interviews.

[5]Two well-networked sources did reach out on my behalf to officials in the Beijing Propaganda Department, and I did establish contact with a high-ranking Shenzhen official who was well-

to some high-ranking executives in major Internet companies, senior newspaper editors, and academics who regularly consulted with officials about 'Internet management', all of which allowed me to partially compensate for the lack of official access.

Clearly, this sample was not random. If the goal were to collect a representative summary of Internet practitioner and scholar views on Internet censorship, this would be an issue. It is not because my purpose was instead a) to ascertain matters of fact related to the functions of various bureaucratic departments regarding censorship, as well as each agency's policy ambit and general reason for existence, and b) to acquire a sample (albeit nonrandom) of informed opinions about leader intentions with respect to Internet control and bureaucratic reform, especially the thinking of elites (roughly, members of the CCP Central Committee and above). Regarding the first objective, bureaucratic purpose is inter-subjective – by definition mutually agreed upon and widely shared among all insiders in a given community. Thus, if several interviewees who were all part of the same community gave similar answers, I was able to draw a reliable inference about the portion of the bureaucracy they interfaced with.

Concerning leader intentions, interviewees' educated speculations were not intended as standalone evidence, but rather to be used alongside a close reading of Internet-relevant Party policy documents. Although inferring individuals' intentions from publicly available documents is fraught with uncertainty, the evidence presented below is still sufficient to establish an intensification in official thinking about reforming the Internet bureaucracy beginning around 2011-12, but one rooted in longstanding ideas about information control. Leaders did not re-invent

connected in the city's tech sector. These individuals all declined to be interviewed after learning my specific topic.

their attitudes regarding the Internet from whole cloth; indeed, the basic objectives of such a regime showed continuity pre- and post-2011. Instead, what changed around 2011 was the intensity and urgency with which leaders sought to reshape the Internet bureaucracy to implement more active management, a development I explore further in the following section.

3.3 Party Leaders: Seizing Social Media's "Commanding Heights"

Former paramount leader Deng Xiaoping's famous dictum "social stability overrides everything" (*shehui wending ya dao yiqie*) has profoundly shaped Chinese leaders' thinking not only about real-world popular mobilization, but also the Internet and social media.[6] Any analysis of how Party elites weigh the costs and benefits of firmly regulating online social spaces must first acknowledge that leaders' concern about these technologies' potential both to spur collective action, and to effect a longer-term change in popular attitudes toward the regime, is a limiting factor in every related decision they make. All groups of interviewees consistently echoed this theme, which also squares with recent quantitative research about online collective action (King, Pan and Roberts, 2013; 2014). One commentator at a Beijing newspaper attributed this depth of leaders' fear to their experiences as victims of persecution from the mobilized masses during the Cultural Revolution, suggesting that both Xi Jinping and Cyberspace Administration of China (CAC) director Lu Wei were especially affected by this horrific past and determined to

[6]Throughout, I refer to the Internet and social media interchangeably. While social media is only one segment of online activity, for purposes of controlling online discourse it exemplifies what officials view as the Internet's most dangerous characteristics.

maintain the Party's grip on communication channels.[7] Officials' precise concern, to paraphrase one Beijing academic, is the 'slippery slope' argument: leaders fear that if they allow speech on certain topics, discussion could veer in a direction much more hostile to the Party's image.[8]

Officials' view of social media's mobilizing potential therefore shaped their interpretation of the state of the Chinese Internet during the 2000s, or as one interviewee put it, ten years of "chaos" (*luan*), a reference that poignantly evokes past periods in CCP history of disorder and breakdown of authority.[9] While such an uncontrolled situation persisted throughout the 2000s, in retrospect elites viewed 2009-12 as particularly disorderly, both in terms of new forms like microblogs spurring actual collective protests, and in terms of more diffuse and longer-term harm to the Party's image resulting from a string of online scandals – food safety issues, local environmental protests, conflicts over land rights, and a host of other issues. While such incidents tended to reflect poorly on officialdom generally and served as an embarrassment to the top leadership, elite-level thinking was not the only justification leaders cited as proof of a 'chaotic' Web; multiple interviewees also emphasized that they believed the public as well as leaders viewed the Internet as 'out of control'.[10]

[7]Interview #48, BJ, 4/16/15.

[8]Interview #2, BJ, 9/10/14. The interviewee did not use the words 'slippery slope'; it is my interpretation of his remarks originally in Mandarin.

[9]Interview #14, BJ, 11/4/14.

[10]Examples interviewees gave, referencing similar speeches by authorities, include so-called "human flesh" searches (*renrou sousuo*), where netizens would use online information to hunt down and expose alleged corrupt officials, effecting a form of vigilante justice; unverified rumors; and the so-called "Internet Water Army" (*wangluo shuijun*) of hired agencies/PR firms enlisted to bolster a client's (or knock down an opponent's) reputation. Interviews: #9, BJ, 9/29/14; #24, BJ, 12/10/14; #28, HK, 1/21/15; #35, SZ, 3/4/15.

3.3.1 Social Media as 'Experiment' (2009-12)

In the subsequent 2013-14 crackdown, leaders attributed responsibility for this situation to two primary groups of actors: the Internet companies themselves, and influential online commentators: celebrities, lawyers and other public figures.[11] Internet companies were held responsible as the ultimate legal responsibility bearers, while bloggers were blamed for spreading malicious and unverified information. While President Xi and other Party elites retroactively decried these actors' lack of discipline, in reality the situation was partially a result of leaders' own deliberate choice to treat China's late 2000s surge in online activity, especially social media, as an experiment. One foreign correspondent who had been stationed in Beijing during this period argued that officials relied on social media as a way to measure public opinion.[12] Another academic interviewee also referred to *Weibo* as 'experiment', while adding that this experiment was "instrumental" rather than reflective of leaders' normative beliefs.[13]

If leaders viewed some liberalized discourse online as instrumentally useful, however, then to what purpose? Especially during the Hu Jintao administration's latter years, reform-minded leaders came to view rising corruption as a major threat. Multiple interviewees mentioned that from leaders' view, one of social media's major benefits was to hold local officials in check by providing bottom-up reporting on corruption, environmental disasters and other problems.[14] In another example, a prominent Shanghai source with strong media official connections, and

[11] Sina.com in particular fell into disfavor with the top leadership after promoting 'hot topics' (*remen huati*) that were often spread by these high-profile bloggers. Interview #21, BJ, 11/27/14.
[12] Interview #36 (via Skype while in Shenzhen), 3/6/15.
[13] Interview #30, HK, 2/3/15.
[14] One especially clear example came from a Chinese tech industry foreign expert. Interview #44, BJ, 4/3/15.

a Beijing news company employee independently suggested that this logic even extended to high-profile cases such as the 2011 Wenzhou train incident, which involved the collision of two high-speed trains and official attempts to suppress media coverage of the disaster. The cover-up failed after bloggers posted images of officials at the scene on *Weibo*, leading to a massive online outcry. Both interviewees claimed that top leaders used online criticism of how the government handled the tragedy to take down former Minister of Railways Liu Zhijun, who was later charged with corruption.[15]

Thus, while officials were clearly concerned about social media's detrimental effects as early as 2009-11, the platform was not entirely without strategic benefit for them during this period. In fact, many interviewees volunteered the idea that in their view, China's leaders were pursuing some variant of 'smart' censorship, restricting both collective action and broader threats to Party legitimacy while allowing some space for targeted criticism.[16] While instances of leaders opening up social media space did not end completely after 2011, the Arab Spring and Wenzhou train accident can nonetheless be identified as turning points that led leaders to adjust their formula toward tighter control.[17] This wake-up call entailed

[15]Interviews: #22, BJ, 12/3/14; #25, SH, 12/13/14. While these interviewees are not regime insiders and cannot know top leaders' intentions for certain, they are representative of relevant outsiders' thinking about the Wenzhou incident.

[16]All interviewees who volunteered an interpretation of 'strategic' or 'smart' censorship *without* me prompting them are cited here (including cases where interviewees did not reference an overall strategy, but used one or more examples to illustrate elites' broader strategic thinking): #4, BJ, 9/6/14; #15, BJ, 11/5/14; #16, BJ, 11/12/14; #18, BJ, 11/16/14; #22, BJ, 12/3/14; #25, BJ, 12/13/14; #35, SZ, 3/4/15; #36, SZ (via Skype), 3/6/15; #37, GZ, 3/9/15; #39, BJ, 3/17/15; #44, BJ, 4/3/15; #45, BJ, 4/8/15. However, a few interviewees did offer non-strategic explanations for the variation in censorship, such as elites' inaction or internal divisions: #17, BJ, 11/13/14; #30, HK, 2/3/15; #43, BJ, 4/1/15.

[17]Although the language of 'turning point' is difficult to falsify, the fact that leaders adopted new language that "Internet development and supervision urgently need to be strengthened and reformed" at the Sixth Plenum, which occurred only months after these events, supports this interpretation (see below footnote). Additionally, three interviewees explicitly mentioned, unprompted, that either the events in the Arab world of early 2011, or Wenzhou were pivotal moments that influenced leaders' thinking. Interviews referencing Arab Spring's role: #14, BJ, 11/4/14; #41, BJ, 3/24/15. Interview referencing Wenzhou incident: #37, GZ, 3/9/15.

leaders' attempts to reconcile two disparate impulses, which were reflected in a concluding statement from the Sixth Plenary Session of the 17th Party Congress in 2011.[18] On the one hand, the need to tighten control over social media became apparent, as leaders admitted they needed to "speed up the formation of an Internet oversight system that combines the force of the law, administrative supervision, industry self-regulation, technical guarantees, public oversight and the education of society" – reforms ostensibly designed to protect user interests and promote a "healthy Internet culture", but also to prevent the emergence of counter-narratives that might threaten the Party's or top leaders' image.[19] On the other hand, in discovering the need to "seize the high ground" in spreading Internet information, Party elites also had a more proactive vision in mind: to "implement the policy of using the Internet in a positive way" and to "strengthen guidance of online public opinion; and promote ideological and cultural themes." One interviewee, a Beijing news editor, offered an eyewitness account, relating how in 2011 he attended a meeting with the editor in chief of People's Daily, who told the assembled editors that they had to be innovative and seize the "ideological battlefield" of social media.[20]

Here, leaders went beyond increasing efforts to restrict the Internet's negative effects, to cultivating a positive image of the Party. One interviewee at a major Beijing technology company attributed this motivation to leaders' sense of lost ideological legitimacy in the reform era, as well as more material concerns like social inequality that threatened the Party's claim to represent all Chinese.[21] Two other

[18] "Decision of the CCP Central Committee on Major Issues Pertaining to Deepening Reform of the Cultural System and Promoting the Great Development and Flourishing of Socialist Culture." Passed at the Sixth Plenary Session of the 17th CPC Central Committee, 10/18/11. Translated by the English Section of the Central Document Translation Department of the Central Compilation and Translation Bureau, Beijing, China. Source: www.cctb.net.

[19] Ibid.

[20] Interview #57, SH, 6/17/16.

[21] Interview #16, BJ, 11/12/14.

interviewees noted President Xi Jinping's emphasis on creating a "positive" online environment; implicitly, filtering out 'negative' speech, much of which criticized the Party or specific leaders.[22] To be sure, elites' conception of "public opinion guidance" as media strategy long predated the Internet: this term has roots in Party leaders' and propaganda officials' efforts to reassert control over the press and establishment media following the 1989 Tiananmen movement. Yet while the concept was not new, the way it had to be operationalized in social media versus older formats was radically different, requiring a far more bottom-up approach to shaping viral discussion spaces like *Weibo* without killing the very dynamism that attracted young, educated demographics to the platform. In short, it required the Party to cultivate its own online commentators in addition to restraining celebrity bloggers.

With this considerable challenge, leaders recognized around 2011 that they were falling short on both negative, and positive means of control. On the negative side, attempts at giving bloggers some space to editorialize about current events while selectively applying censorship had failed in the eyes of many elites.[23] The bureaucracy responsible for enforcing censorship was fragmented, with local Public Security Bureaus – which are decentralized actors under the direction of municipal governments and districts – making judgment calls regarding the Party's (or just as often, petty individual) interests that went far beyond the 'Internet Police' (*wangjing*) mandate, according to a former editor at a major central Party newspaper.[24] Concerning positive control, the Party faced still greater institutional

[22]Interviews: #27, HK, 1/16/15; #28, HK, 1/21/15. While President Xi has emphasized such "positivity" to a greater extent than his predecessor, the idea was firmly entrenched as early as 2011 during Hu's last years; the word "positive" appears ten times in the Central Committee's 17th Congress 6th Plenum statement.

[23]Interview #44, BJ, 4/3/15.

[24]Interview #47, BJ, 4/14/15. Regarding 'individual' interests, a form of corruption involving Internet company employees accepting money, or being pressured from unauthorized people to delete posts their clients found 'undesirable' was also a major impediment to top leader attempts

weakness in the inability of the propaganda system to adapt to new media. Several interviewees, particularly journalists and editors who regularly received orders from propaganda officials, noted that the CPD and its provincial-level counterparts suffered from numerous weaknesses that were particularly detrimental in the Internet age, such as being slow in reacting to breaking incidents,[25] and failing to grasp social media's importance in reaching new audiences; this last point, two interviewees noted, was attributable to officials' "old" age.[26] Additionally, although propaganda officials did sometimes grasp the need to extend outright bans on topics (as they often have for press coverage) to social media, one interviewee who regularly monitors the implementation of online censorship told me he had found instances where such directives were flouted online even as traditional media publications complied. While the CPD, as a leading Party organ, theoretically could enforce its will upon all media, its ability to do so *de facto* on the Internet was seriously in question.[27]

A host of problems concerning what leaders perceived as an out-of-control Internet thus factored into their resolve to tighten control while preserving 'smart' censorship's most useful aspects. In attempting to do so, leaders found the existing central bureaucracy inadequate to the task. All that said and despite numerous weaknesses, China's leaders started efforts to strengthen the censorship system with important assets not available to other authoritarian states. One such asset was the presence of vibrant domestic Internet companies was an important prerequisite for leaders' success, which the next section considers.

to regulate online space. Interviews: #44, BJ, 4/3/15; #15, BJ, 11/5/14.

[25]Interview #10, BJ, 10/2/14.

[26]Interviews: #23, BJ, 12/5/14; #49, BJ, 4/22/15.

[27]Interview #16, BJ, 11/12/14.

3.4 Internet Companies' Symbiotic Relation to State Authority

The vibrancy of China's Internet industry contrasts sharply with many of its autocratic peers. While explaining the tech industry's rapid development is an economics or business topic for industries in advanced democracies, in China the sector's abrupt rise constitutes a political puzzle given continuing heavy state involvement in the market. How has such a dynamic sector come to exist in China, particularly since the state retains substantial ownership in television and newspapers? I argue that several factors that long predate the events of 2011 explain the Chinese tech industry's success despite stringent regulation, and its ability to form the scaffolding upon which leaders could carry out a sophisticated censorship strategy. One could begin with obvious economic and cultural factors: China's large and increasingly affluent population, high Internet adoption rates, and the usage of Chinese characters as a common written language (and walling off the sinophone world from more globally mixed language regions). Yet equally important has been the Chinese government's investment in the IT industry, notably the establishment of 'technology parks' for research and development like Beijing's Zhongguancun district. Such investment has not been merely a matter of national policy, but of top leaders' personal interests; as an example, according to a leaked Beijing U.S. Embassy cable, Hu Jintao's son-in-law "ran" Sina.com.[28]

Although Party investment in Internet media is certainly in part for financial and economic reasons, Party leaders are increasingly doing so in order to practice a form of censorship long prevalent in the West: editorial control through owner-

[28]Source: Leaked U.S. Embassy Beijing diplomatic cable. July 9, 2009. See `http://www.wikileaks.org/cable/2009/07/09BEIJING2112.html`.

ship. According to a senior figure at a privately held media company, leaders have awakened to the fact that direct ownership is a very effective means of control.[29] Internet media companies, for their part, often depend on large infusions of external financing to stay afloat as they struggle to monetize online services. While some less news content-oriented companies like Tencent and e-commerce sites like Alibaba have been able to monetize a range of services on their platforms, the situation is very different for microblogs like Sina *Weibo* and news portals. My interviewee explained that *Weibo* in particular was very expensive to maintain (in terms of technology and software developer costs), and as the government had a vested interest in shaping the platform's content, it became a natural investor to which Sina executives were then beholden.[30] While lack of profitability has been a serious threat to microblogs elsewhere – e.g. Twitter in the U.S. – in China government-directed investment has kept these services afloat while ensuring their parent companies' political loyalty.

Third, while censorship regulations have been onerous and "a major time suck" for company executives according to one domestic company source,[31] it would be a vast overstatement to assert that they have crippled the sector. One company official who was responsible for implementing government censorship directives bluntly stated that the cost of carrying out censorship was simply "not enough to matter", mentioning that the company only needed one or two full time employees for this task.[32] Internet companies also provide surveillance and intelligence information on citizens.[33] On the positive side, the Chinese tech sector has been

[29]Interview #54, BJ, 6/8/16.
[30]Interview #54, BJ, 6/8/16.
[31]Interview #1, BJ, 9/9/14.
[32]The individual was referring to search engine censorship, which is indeed much less labor-intensive compared with microblogs and other online spaces. However, the interviewee clearly intended to make a broader point about Internet censorship overall. Interview # 53, BJ, 6/8/16.
[33]One company insider mentioned how Baidu supplied search data about the Falun Gong to the government in 2004. He said Baidu had proven similarly useful to the government in other

beneficial to the national economy, with companies like Tencent serving as market 'disruptors' by integrating services ranging from digital payment to taxi hailing into their platforms.[34] Based on numerous interviews and contrary to perceptions in the West, it is simply not true that the censorship burden has stifled Internet company innovativeness, including in the online media sector. Except for editorial content limitations and the need to filter or delete some user-generated content, companies are free to attract clicks and views however they see fit.

A fourth factor is that the same censorship requirements that impose a limited burden on Internet companies also offer them protection from foreign competition; as domestic companies become more compliant with censorship directives, they remain acceptable to Chinese leaders while foreign companies struggle with both market entry, and complying with directives once in-country.[35] However, despite their privileged position in China's economy, Internet giants' freedom to innovate and make money is still not entirely safe from government meddling, as some central-level officials have a stronger interest in 'the market' over the Party's political goals than others.[36] Company executives thus expend great effort to ensure they remain in the good graces of relevant agencies.

In sum, China's large Internet companies and the state enjoy a symbiotic relationship where the former enjoys much-needed state investment and market protectionism, while granting the latter compliance with censorship directives, and even proactively working to exercise "self-discipline" in ensuring the spread of pro-Party

cases. Interview #47, BJ, 4/14/15.

[34]One Sina employee called China's Internet sector the "most innovative in the world" except for politically sensitive content. Interview #51, BJ, 6/6/16.

[35]One senior Internet company representative, although stating that censorship's primary intent was not protectionism, nonetheless admitted that it had that benefit. Interview #39, BJ, 3/17/15.

[36]One interviewee at a major Chinese media company bluntly stated that the State Administration of Press, Publications, Radio, Film and Television (SAPPRFT) "doesn't give a **** about the market." Interview #22, BJ, 12/3/14.

messages online. While the existence of this symbiosis does much to account for robust state control over Internet companies, the factors in this section were already either strongly present or at least underway prior to Xi taking power in 2012, and so cannot explain the bureaucracy's increasingly robust control after that date. Doing so requires a look at the specific state agencies and actors to which companies are *most directly* accountable and how such a bureaucratic configuration has changed in the Xi era.

3.5 The Internet Bureaucracy Pre-Reform (1990s-2011): Partial Fragmentation

In reforming the Internet bureaucracy, Chinese leaders did not begin from scratch prior to the events of 2011. To the contrary, a handful of agencies 'held down the fort', enabling top leaders to achieve their minimum objective during urgent online breaking events: to effectively suppress and delete information they perceived as harmful to the Party's or their personal interests. This section analyzes these pre-existing agencies beginning at the municipal level.

3.5.1 Holding Down the Fort: Actors At the Provincial/municipal Level

An initial key aspect to understanding China's Internet bureaucracy is that it is a two-tiered system: censorship directives can and do come either from the central

government, or from the provincial level, while major policy decisions are made centrally.[37] While such decentralization often leads to bureaucratic fragmentation and conflict between levels in other policy areas, in regulating China's Internet giants the situation is greatly simplified by the fact that most major companies are located in Beijing, with a few in Shanghai and Guangdong, and almost none anywhere else.[38] Such a situation contrasts markedly to other economic sectors in China, where production occurs in multiple jurisdictions. The fact that the number of lower-level governments is minimal allows the center to both delegate much oversight to these few local governments, and intervene expediently when needed. The following sub-sections discuss the essential features of the most important local and central actors charged with Internet regulation.

The Public Security Bureau (e.g. "Internet Police")

The Beijing Public Security Bureau (PSB, a.k.a. "the police") play a vital role as the enforcers of both written Internet laws and regulations, and the political will of Party elites. While the Beijing PSB is nominally affiliated with the central Ministry of Public Security, in fact it is under the direct leadership of the Beijing municipal government, from whom it receives its budget and personnel. The Beijing police, like all local police throughout China, are thus decentralized, dependent on government authorities in the jurisdiction where they are based rather than on higher-level public security officials. Within the Beijing PSB there is an Internet unit, popularly known as the *wangjing* (literally "Internet police"). Due to China's system of localized media control, social media sites registered in Beijing are thus under the Beijing *wangjing's* direct oversight – in fact, one interviewee with exten-

[37] In the Chinese system, Beijing Municipality is the administrative equivalent to a "province".
[38] To simplify, in the following analysis I assume that an example company is located in Beijing.

sive contacts in the Beijing technology industry noted that major companies like Sina have Internet police "in-house" that are constantly monitoring user posts.[39]

This decentralized situation, even prior to reform, did not preclude the PSB expediently enforcing 'priority' censorship orders from the center as well as Beijing municipality during urgent breaking events, but it did result in a lack of clarity regarding the appropriate scope of *wangjing* activities, and cooperation with other units in top-level initiatives to solidify Internet control. The police's role in implementing higher-level censorship policies is important because they are the main agency with day-to-day enforcement capacity. Long before 2011, central authorities began pushing legal reform in an effort to clarify the functions of law enforcement, including online. A typical example was an amendment to the 2010 Law on Guarding State Secrets, which contained new provisions specifying how Internet companies were to cooperate with the PSB in the investigation and handling of state security leaks.[40] However, the police's greatest strength – their ability to promote anti-crime and "national security" interests in Internet management – was also a major limitation pre-2011; the PSB then had (and still has) no financial or interest-based stake in regulating online space because taking responsibility for more politicized censorship decisions would do nothing to increase their budget or personnel.[41] Nonetheless, in the absence of clear superior authority to decide what social media 'hot topics' were 'politically sensitive', prior to the reforms begun in 2011 such judgment calls often ended up in the Internet police's hands. According to a tech sector worker, the Internet companies "dare not" disobey the PSB even though "its authority is limited to security matters." He noted that the police "don't have the right" to censor politically sensitive content,

[39]Interview #44, BJ, 4/3/15.
[40]Source: `http://www.gov.cn/flfg/2010-04/30/content_1596420.htm`.
[41]Interview #22, BJ, 12/3/14.

but "do it anyway."[42] The Beijing PSB's *de facto* political power as regulator of Sina *Weibo* and other major services also caused cross-jurisdictional conflicts, as officials or police in other provinces would have to lobby Beijing officers to order companies to delete unwanted content.[43] Such fragmentation was a major target of the post-2011 restructuring.

The Beijing Internet Propaganda Culture Management Office/Beijing Internet Information Office (a.k.a. "Internet Management Office")

The other pivotal office overseeing Internet censorship in Beijing goes by three different names. For foreign English speakers, it is referred to as the "Beijing Internet Management Office", a title that aptly reflects its broad functional role. In Chinese, it has two names. Prior to 2013, it was little publicized and known to insiders as the "Beijing Internet Propaganda Culture Management Office" (*Beijingshi hulianwang xuanchuan wenhua guanli bangongshi*),[44] a title that reflects its position in China's propaganda system. Before 2013, it was a *party*, not governmental body under the leadership authority (*lingdao guanxi*) of the Beijing Municipal Propaganda Department.[45] In 2013, this office was given an official *governmental* name – the "Beijing Internet Information Office" (*beijingshi hulianwang xinxi bangongshi*) – and was tasked with undertaking a "professional consultative"

[42]Interview #20, BJ, 11/20/14.

[43]Interview #9, BJ, 9/29/14. The interviewee's specific statement was that other jurisdictions had to lobby the Beijing "city government." However, the Beijing PSB would be the ultimate target of such a lobbying effort.

[44]Insiders also refer to it as the *wang guan ban*, literally "Internet Management Office" for short, an abbreviation that directly matches its English name. Foreign reports have continued to refer to it as the "Internet Management Office" even after its 2013 Chinese name change.

[45]The Beijing Municipal Propaganda Department, which itself is under the Beijing Communist Party leadership, still holds direct authority over the Internet Management Office post-reform. However, the newly-empowered central-level Cyberspace Administration (CAC) exerts much greater influence than its predecessor, a situation I discuss below.

role (*yewu guanxi*) with a host of other municipal-level agencies that deal with Internet regulation – including the Beijing Internet police. Thus, this office now fuses Communist Party, and Beijing government authority under one roof, a situation referred to in Chinese as *yi men hang, liang kuai paizi* or "one door, two signboards."

Regardless of its name, this office is *the* office directly responsible for issuing orders to the Internet giants in Beijing to delete unwanted content.[46] Its authority to order deletions far exceeds the PSB's; while the police generally directly give deletion orders only on 'security' or crime-related matters, the Internet Management Office often does so for unwanted content that in its (or its superiors') judgment a) threatens social stability, b) harms the Party's image or agenda, or c) insults or even comments on top leaders' activities, to name just the most common examples. As a part of China's powerful propaganda system in the key jurisdiction of Beijing, the Internet Management Office is very powerful, despite the fact that its formal rank is as a *ban* or "office", a lower-ranking (and typically, smaller and less well-resourced) unit compared with the PSB, which is a municipal "bureau" (*ju*).[47] The reason has to do with the propaganda system's exalted role within Chinese governance. Not only is the "Party above the government" — in China, the Communist Party's organizations set the general political line, while "government" agencies administer and implement this line — but the Propaganda Departments at various levels are among the most important of all Party organs, given the CCP's longstanding emphasis on propaganda and ideology. This means that the Beijing police are unlikely to take any Internet enforcement action that

[46]Interviews: #9, BJ, 9/29/14; #11, BJ, 10/14/14; #14, BJ, 11/4/14; #21, BJ, 11/27/14.

[47]Interview #42, BJ, 3/24/15. As an example of this office's power, it was the body that sent out the directive to Internet companies in March, 2015 ordering Web portals to remove the controversial air pollution documentary *Under the Dome*. I discuss this incident below. See `http://chinadigitaltimes.net/2015/03/minitrue-clamping-dome/`.

would contradict either the political will of the Beijing, or the central propaganda authorities.

While the Internet Management Office enjoyed clear strengths as a "one-stop shop" for political Internet censorship decisions in Beijing, it also suffered from serious limitations prior to the post-2011 reforms. First, it lacked formal "consultative relations" with its functional administrative counterpart at the central level until the CAC was established. Second, the office's authority, through broad in principle as a Party body, was limited by the fact that it did not have formally defined relation to the Beijing Internet Police or other municipal-level "relevant agencies." Addressing these deficits was a major task of reforms begun under Hu, and greatly accelerated under Xi.

3.5.2 Division at the Top: the SCIO/SIIO, and Propaganda Department

Perhaps due to the necessity of interfacing with the booming Internet sector in China's capital, the resources of Beijing municipal actors outstripped equivalent capabilities at the central level. Until 2011 (and arguably, until 2013), the central *government* (not Party) lacked any administrative analog for the Internet and social media to the CPD's broad role in regulating newspapers. Nevertheless, a designated administrator in charge of regulating "Internet content" did exist: the State Council Information Office (SCIO), a.k.a. Office of Foreign Propaganda (OFP).

The SCIO/OFP

In contrast to the well-defined roles of the Internet Management Office and Internet police at the Beijing municipal level, leaders initially placed central authority over regulating "Internet content" in the hands of the OFP, which is "one and the same" with the SCIO (Brady, 2008).[48] Although the OFP's primary mandate is foreign propaganda, the Internet was still put under its portfolio despite the fact that the Chinese Internet is heavily domestically oriented (Chinese netizens primarily visit domestic websites). However, this awkward situation was ameliorated by the establishment of an Internet Affairs Bureau within OFP/SCIO to specifically monitor Internet content. While OFP/SCIO and its Internet bureau had enormous authority under the State Council's direct leadership, like the Beijing Internet Management Office it suffered from the drawback that its formal responsibilities and oversight relation to other central-level agencies were poorly defined. Nonetheless, the OFP/SCIO would frequently send out both broad Internet policy directives, and specific censorship bans on matters of national importance, while leaving to lower-level authorities less critical 'hot topics' or more specific follow-up instructions.[49]

In 2011, the Internet bureau of the OFP/SCIO was broken off into a new agency, the State Internet Information Office (SIIO). This office, later given expanded authority as the Cyberspace Administration of China (CAC), built on the bureaucratic lineage of OFP/SCIO to become the linchpin of re-centralized censorship. Before considering the SIIO/CAC's post-reform powers, however, the next section examines one final actor in its pre-reform state: the Central Propaganda

[48]This was another instance of *yi men hang, liang kuai paizi.*
[49]Interview #9, BJ, 9/29/14.

Department (CPD).

The Pre-reform Central Propaganda Department (CPD)

As the Party's key media control institution, the Central Propaganda Department might be expected to be leading the charge to "seize the commanding heights" of social media, as the CPD has with newspapers, radio and TV. On this topic, several interviewees consistently repeated two points: 1) top-level propaganda officials and the Party leadership were enthusiastically committed to using social media, but 2) they were "behind", out of touch, or lacked Internet experts.[50] Indeed, respondents cited a host of issues with the CPD's approach to the Internet prior to (and even during) reform. One explained that in his view, a major problem was the Department's persistence in applying traditional 'broadcasting' propaganda techniques to the Internet, even though it is a more user-centric medium.[51] Another issue was response speed; the CPD simply "couldn't keep up" during *Weibo's* first two years (2009-10), a time in which the pace of stories broken via the Internet accelerated rapidly.[52]

While such issues certainly affected the CPD's ability to adapt, a larger barrier was structural: as a Party rather than administrative/government body, the CPD has no direct regulatory authority over Internet companies.[53] This matters because although the Department's clout with companies is enormous, the CPD does not (and likely cannot) micromanage the major Internet companies; it is used to having its orders obeyed with print media and not very good at 'following up' on deletion

[50]Interviews: #10, BJ, 10/2/14; #20, BJ, 11/20/14; #23, BJ, 12/5/14; #48, BJ, 4/16/15; #49, BJ, 4/22/15.
[51]Interview #12, BJ, 10/16/14.
[52]Interview #36, (via Skype while in Shenzhen), 3/6/15.
[53]Interview #15, BJ, 11/5/14.

requests in the much more chaotic environment of social media. Even before 2011, the CPD had officials who concurrently held government posts in agencies, like OFP/SCIO, that could issue clear, binding orders and had the resources to monitor their implementation. Thus, while the Department could often indirectly influence Internet censorship, it had to rely on intermediaries.[54] Although this partly reflects the principle that the CPD should not duplicate other state agencies' regulatory functions (Brady, 2008), it may also reflect the fact that the CPD is simply not well suited to managing Internet content.[55]

The CPD's lack of direct action contrasts sharply with the Beijing Internet Management Office. The latter's local-level innovativeness became especially apparent under the tenure of Lu Wei, who as head of the Beijing Propaganda Department oversaw both the Internet Management Office's development, and the enlisting, according to Lu's own statement, of "60,000" Internet propaganda workers on the Beijing government's payroll and "two million" employed in propaganda off-payroll.[56] Perhaps not coincidentally, Xi Jinping picked Lu in 2013 to head the CAC and to spearhead Internet regulatory reform.

[54]One high-ranking Internet company employee who dealt with government censors noted that in all his years, he had never received an order from the CPD. Interview #16, BJ, 11/12/14. Also relevant is Interview #22, BJ, 12/3/14.

[55]Interviews: #16, BJ, 11/12/14; #31, HK, 2/4/15; #48, BJ, 4/16/15.

[56]On its face, this number seems fantastic as it implies that roughly one out of every ten Beijing residents (city population 20 million) is engaged in online propaganda work. However, the South China Morning Post claimed to verify this number with a call to Beijing Internet Information Office. Lu gave the figure at a "conference attended by propaganda department heads in the city" on January 17, 2013. See `http://www.scmp.com/news/china/article/1131287/about-10pc-beijing-residents-work-propaganda-services`. Additionally, one interviewee arrived at a similar number by explaining that Lu designated 4-5 propaganda liaisons within each *shiye jigou* (city services unit) in the Beijing City government, of which there are around 20,000. This would put the total at around 80,000-100,000, close to Lu's first figure. Interview #57, SH, 6/17/16.

3.5.3 Analysis: Adequately Reactive, Inadequately Proactive

The above descriptions represent the state of China's censorship system in early 2011. Figure 3.1 depicts the authority relations among this system's various components. While this system was far simpler than bureaucratic structures in other policy areas (see Mertha 2005; 2008), companies were still answerable to multiple entities for both discrete censorship orders and broader policy. For example, the Beijing Internet Management Office, and the Beijing Internet Police (PSB) could both issue orders for companies to delete content – yet neither reported directly to the other, and while the OFP/SCIO outranked these municipal-level actors, pre-reform it did not have formal *yewu guanxi* with either one.[57] This fragmentation made life more complicated for Internet companies in deciding whose orders to follow: one company insider characterized the situation as "a mess".[58] Another Vice-President level insider who dealt directly with censorship described a system in constant flux that "changed every few months."[59] Still another consequence of fragmentation was to increase opportunities for corruption, as local officials fearing online exposure would pay Internet company employees to delete posts.[60]

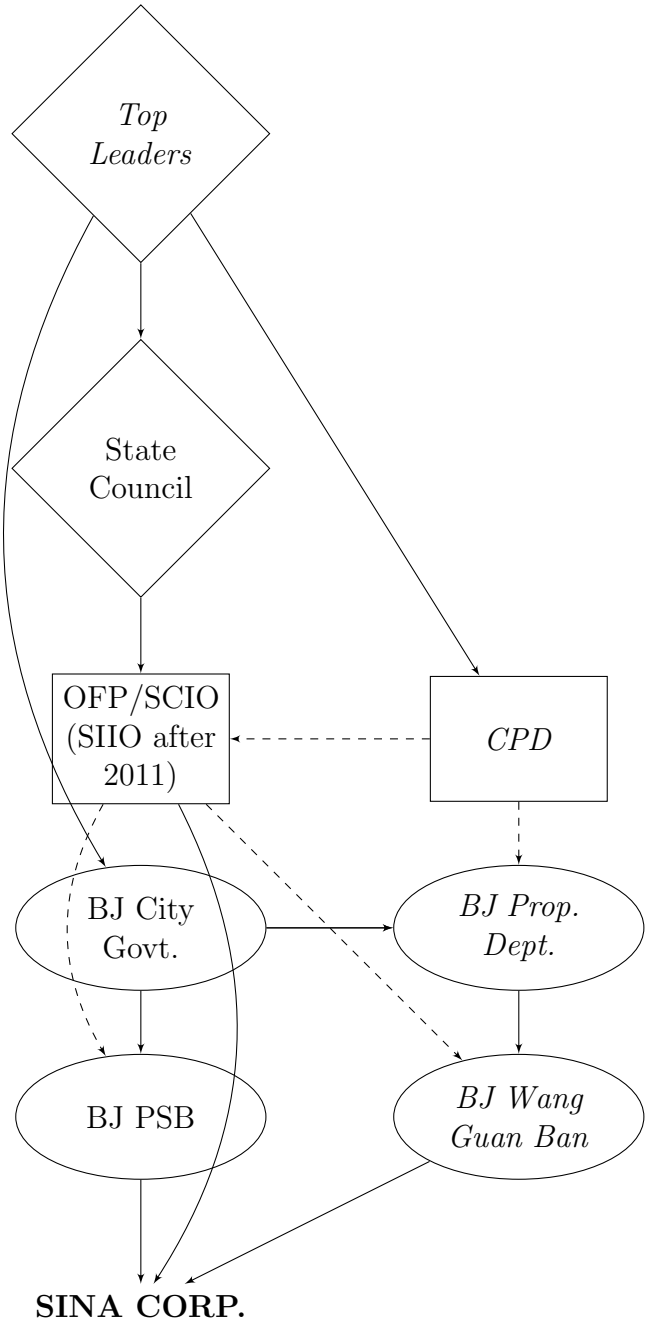
[57]For simplicity's sake, the schematic excludes other somewhat-relevant actors such as the Culture Ministry, the Ministry of Industry and Information Technology (MIIT) and the State Administration of Press, Publications, Radio, Film and Television (SAPPRFT). These actors matter in regulating particular aspects of the Internet. However, I did not include them in the analysis due to their minimal roles in day-to-day regulation of online blog posts and news articles which are the subject of this analysis.

[58]Interview #39, BJ, 3/17/15.

[59]Interview #7, BJ, 9/25/14. The interviewee made this comment in September, 2014, suggesting that even after the reforms, some inter-bureaucratic conflict remains. Of course, this does not mean that no effective streamlining has taken place.

[60]Interview #44, BJ, 4/3/15.

Figure 3.1: Fragmented Authority: The Chinese Social Media Censorship System Prior to Reform



Note: Diamond = leadership pinnacle; Rectangle = central level government agencies and Party organs; Oval = provincial (Beijing municipal) level. Solid lines = binding authority (*lingdao guanxi*); dashed lines = strong influence (*yewu guanxi*), but no binding authority. Regular text = government; Italics = Party organ. **Exception**: The solid lines pointing to Sina Corp. are not technically *lingdao guanxi*, since these can only exist between bureaucratic entities. But multiple interviewees indicated that Internet companies including Sina obey orders from the actors I have shown in this chart as totally binding, just as is the case for *lingdao guanxi*.

Yet while the system suffered from numerous weaknesses, it was still robust enough that on priority topics, central-level officials or their Beijing-level counterparts could still order Sina and other companies to delete posts within minutes or hours. The system thus worked partly through redundancy – on some level it did not matter which entity issued the order as long as companies obeyed it. This system was very good at reacting to undesired news or trending topics for two reasons: first, the number of major companies to regulate was small and they had clear incentives to comply, and second, the right agencies – especially the Beijing Internet Management Office, with support from the Internet police – were in place to give, monitor, and follow up on orders. However, due to fragmentation and a lack of central leadership, the system was poor in two other aspects: maintaining censorship discipline during day-to-day (non-emergency) events, and combining censorship with positive propaganda. The turning point of 2011-12 then laid bare this incapacity and provided momentum for further reform.

3.6 Reform and Restructuring (2011-)

Leaders' efforts at reform did not coalesce immediately after the Wenzhou incident. Rather, most major reforms had to await completion of the 18th Party Congress in November, 2012 and the transition to Xi's leadership. One notable exception was the upgrading in rank of the SCIO Internet Affairs Bureau to become a separate office reporting directly to the State Council: the State Internet Information Office, or SIIO (*guojia hulianwang xinxi bangongshi*), in May 2011. While such a move gave the former bureau increased prestige and autonomy, this step still fell short of establishing a true "Internet czar" to oversee China's Internet-

relevant ministries; the Xinhua news release indicated that the new office would direct "online content management", "oversee government propaganda", and listed several other responsibilities (e.g. the very tasks that were then scattered across other ministries and agencies). The announcement left unclear whether the SIIO would have leadership relations with these to-be-subordinated ministries, or only consultative relations, which would mean the SIIO could not issue binding orders to them.[61]

Additionally, leaders made sporadic attempts at actually implementing long-discussed policy initiatives even before the Congress, using existing structures. In December, 2011, the Beijing PSB, Internet Information Office (Internet Management Office), and the Beijing branch of China's Ministry of Industry and Information Technology jointly announced that they were ordering companies with microblogs registered in Beijing to require users to register under their real names – information that would be checked against police databases.[62] The order also included rules intended to enforce language in the 2010 State Secrets law on "posting and duplicating illegal content, including information that leaks state secrets, damages national security and interests, [or] instigates ethnic resentment, discrimination or illegal rallies that disrupt social order."[63] By April 2012, however, authorities ceased attempting to implement the new rule after heavy pushback from companies.

The above two examples illustrate that the challenges facing the Party's attempts to co-opt rather than crudely suppress social media were not a question

[61]Source: The New York Times. 5/4/2011. "China creates new agency for patrolling the Internet." http://www.nytimes.com/2011/05/05/world/asia/05china.html

[62]Source: The New York Times. 12/16/2011. "Beijing imposes new rules on social networking sites." http://www.nytimes.com/2011/12/17/world/asia/beijing-imposes-new-rules-on-social-networking-sites.html?ref=technology

[63]Source: Xinhua. 12/16/2011. http://news.xinhuanet.com/english/china/201112/16/c_131310381.htm

of intent. Rather, they were a product of a lack of strong central authority and inter-bureaucratic coordination. After the 18th Party Congress, the leadership under Xi addressed this with a two-step maneuver: a) creating a new high-level Party group for overseeing Internet policy and linking it to an elevated SIIO, and b) marginalizing central-level propaganda officials, especially the CPD.

3.6.1 China's "Internet Czar": the Central Leading Group for Internet Security and Informatization and Cyberspace Administration of China

Since major reforms in 2013-14, the Cyberspace Administration of China (CAC) – which is the English name for a joint party/state organ variously referred to as the General Office of the Central Leading Group for Internet Security and Informatization (*zhongyang wangluo anquan he xinxihua lingdao xiaozu bangongshi*) and the State Internet Information Office (SIIO, see above) – has become the undisputed "head honcho" of Internet regulatory organs at the central level. As is evident from retaining the SIIO label, the office is a direct continuation of the SIIO established in 2011. Through its association as the General Office of a form of supra-bureaucratic oversight committees called "leadership small groups" (*lingdao xiaozu*) used by the top leadership to exert control over all ministries, the CAC now unambiguously outranks a host of subordinate ministries involved in Internet regulation, and all equivalent municipal/provincial level bodies, including in Beijing. That is, it is truly *national* in scope. As is the case with similar party/state central level organs, part of the CAC's power stems precisely from its dual status.[64]

[64]Yet another example of *yi men hang, liang kuai paizi.*

As the officially designated state organ in charge of coordinating and where necessary, ordering around ministries such as the Ministry of Public Security (more specifically, municipal-level "Internet police"), the CAC enjoys broad authority to set Internet policy under the direction of its leadership small group. Its responsibilities are sweeping and include regulating Internet content, e-commerce, e-finance, cybersecurity and encryption, and combating online crime, rumors, and pornography. Prior to the CAC's establishment, at the central level nearly all of these policy areas had been claimed by other ministries; for example, the MIIT and PSBs had laid claim to cybersecurity issues, while the Ministry of Culture claimed to be in charge of online anti-pornography campaigns. These ministries are still broadly represented in the new leadership small group, which has representation for nearly all policy areas remotely associated with cyberspace. This leading group was established about a year into Xi Jinping's term, in November, 2013, a key session in which the new leadership announced wide-sweeping reform plans in numerous policy areas. Both the group, and its general office can thus be viewed as Xi's attempt to re-centralize authority over a relatively new and evolving sphere, the Internet, for which the new leadership viewed the existing ministry division of labor as muddled and inadequate.

Both substantively and formally, the CAC differs from existing Internet regulatory agencies. It has been described by various reports as having a "start-up" culture in which employees are among the central government's most likely to "work overtime." It also has "one of the youngest average employee ages of any central government agency, at 37.8 years."[65] Prior to July 2016, its head was Lu Wei, who is not a Politburo or even Central Committee member – a curious lack of rank for the head of such a powerful new agency. Lu's background instead reflects

[65]Source: Council on Foreign Relations Net Politics blog. http://blogs.cfr.org/cyber/2016/07/13/leadership-change-at-chinese-internet-regulator/.

the combination of political reliability, industry knowledge, and policy expertise. The first characteristic is evident from his many years at Xinhua News Agency, while the latter two could stem from his time overseeing the Beijing Propaganda Department, and therefore frequent interactions with Beijing Internet giants. Indeed, various interviewees emphasized both aspects of Lu's background: he is "a propaganda guy",[66] but also "very savvy" and has been willing to meet with tech company illuminati ranging from famous entrepreneur and microblogger Pan Shiyi, to Facebook's Mark Zuckerberg.[67]

Lu's somewhat unconventional background for an official having attained his current rank belied his informal influence as CAC head: he frequently reported directly to President Xi.[68] His three titles during his tenure shed further light on the CAC's dual party/government nature – one observer listed them, "in order of importance", as 1) Vice Director of Propaganda, 2) Head of the General Office of the Central Leading Group for Internet Security and Informatization, and 3) Director of the SIIO.[69] The first title shows that during his tenure, Lu was formally integrated into the CPD, and propaganda system generally. However, in an unexpected twist, Lu was replaced in June 2016 by CAC Vice-Director Xu Lin, who is considered a 'rising star' and had previously served on Shanghai's municipal Standing Committee while Xi was Party Secretary there.[70] While the reasons for the switch remain unclear, Xu (like Lu) is viewed to fit two criteria believed to be Xi's priorities: political loyalty, and a talent for innovative online propaganda.

[66]Interview #44, BJ, 4/3/15.

[67]Pan Shiyi is CEO of SOHO China, and an outspoken public figure on social and environmental issues.

[68]Statement by Sunxian Tang at Workshop #80 of the 2014 Internet Governance Forum in Istanbul, Turkey. Later substantiated by Interview #54, who said that Lu reported "once weekly" to Xi.

[69]Interview #44, BJ, 4/3/15.

[70]It would be premature to assume that Xi was unsatisfied with Lu's performance or that his high-level career is over, as Lu retained his title as Vice Director of the CPD.

Whether Xu will continue Lu's proactive engagement style remains to be seen.

Regardless of who heads it, the CAC on paper is clearly a powerful regulatory body. But what about in practice? How successful has the CAC been both in enforcing its will over other ministries and the Internet giants? On this point, while interviewee responses varied, overall they left little doubt that the CAC has truly become China's "Internet czar", answerable only to Xi himself.[71] Some interviewees did clarify, however, that the CAC was not meant to supersede the functions of existing ministries, but rather to serve as a coordinating body and final authority.[72] The CAC has also not displaced the role of the Beijing Internet Management Office in issuing the most censorship orders to Beijing companies; the center delegates day to day management to the Beijing leadership and the propaganda authorities that serve under them, although the CAC doubtless retains residual influence at the municipal level given that many of its staff were formerly city propaganda officials.[73]

Nonetheless, the CAC has helped Party leaders to centralize the bureaucracy.[74] To some extent, this has in fact meant the transfer of responsibilities for monitoring censorable topics and being the one to give Internet companies the order. One striking example concerns so-called "collective mass incidents" (*qunti shijian*). While King, Pan and Roberts (2013) identified the Internet police as responsible for censorship implementation (p. 1), one interviewee who was a high-ranking editor at a Party newspaper told me that on mass incidents it was the CAC that actually issued the order, saying that the PSB's authority was now limited to nar-

[71]Interviews: #16, BJ, 11/12/14; #20, BJ, 11/20/14; #39, BJ, 3/17/15; #44, BJ, 4/3/15; #47, BJ, 4/14/15. A follow-up trip in June 2016 was more conclusive, with a high-profile interviewee (#54) describing Lu Wei as "the king's man" and stating that he reported weekly to Xi. The interview occurred prior to Lu's replacement by then Vice-Director Xu Lin.

[72]Interview #22, BJ, 12/3/14.

[73]Interview #37, GZ, 3/9/15.

[74]Interviews: #2, BJ, 9/10/14; #44, BJ, 4/3/15.

rower security matters.[75] Such an observation would be consistent with top leaders' growing concern about online collective action, particularly on microblogs, and a desire to re-centralize related censorship decisions. Finally, the CAC has largely replaced the PSB (and its central-level functional equivalent, the Ministry of Public Security) in a range of Internet supervision roles although the latter retains a "day to day" enforcement function.[76] And according to one source the practice of other PSBs calling the Beijing police to ask them to order Internet companies to remove undesired content has ended.[77]

3.6.2 The Marginalization of the Central Propaganda Department

The CAC's attempts to assert control have not come without struggle against other agencies. In particular, multiple interviewees cited examples of tensions that exist between the CAC and CPD. One interviewee interpreted this clash as Xi's attempt (as he has done elsewhere in the bureaucracy) to place his own people within the CPD.[78] Another former Beijing journalist noted that Xi "was not very satisfied" with the CPD's lack of adaptation to new media, and pointed to a recent publicity stunt of Xi being made to visit a local dumpling shop in person and pay with cash himself as the sort of social media-savvy maneuver backed by Xi's people but opposed (to that journalist's knowledge) by the CPD.[79] Another interviewee viewed this conflict in terms of factions, with Jiang Zemin and Liu Yunshan having

[75]Interview #20, BJ, 11/20/14.
[76]Interview #56, BJ, 6/14/16
[77]Interview #56, BJ, 6/14/16
[78]Interview #41, BJ, 3/24/15
[79]Interviews (same subject): #21, BJ, 11/27/14; #57, SH, 6/17/16

backed current director Liu Qibao and other CPD officials' careers.[80] Still another considered Xi's elevation of outsider Lu Wei to have set up a clash between the CAC and CPD.[81]

Unfortunately, given the opaqueness of the process and the recency of still-unfolding reform efforts, we have no way of confirming the exact degree of tensions that exist between the CAC and CPD, but one interviewee did relate convincingly that the former (specifically, Lu Wei) has had his way on at least one important occasion: the decision to allow the air pollution documentary *Under the Dome* to be aired online at a politically sensitive time just before China's National People's Congress in late February 2015. The film caused a political stir and hundreds of millions of views as several online video sites promoted it, but was censored after only one week. My source claimed that Lu Wei personally viewed the film prior to granting permission and supported it, with the CPD in opposition.[82] Lu won out, and the film was allowed to be shown until public commentary about the documentary began to stray far beyond the issue of air pollution and (in leaders' eyes) into more dangerous territory. That the film was aired at all could be viewed as a victory for Lu, although the CPD may have gained support after online discussion got out of bounds. However, in one final piece of evidence supporting Xi's alleged opposition to the CPD, it was chastised by the Central Commission for Discipline Inspection, Xi's signature tool of his anti-corruption campaign, for "weak points like new media."[83] While this criticism could be interpreted as part of Xi's overall attempt to ensure political loyalty by requiring officials to demonstrate

[80]Interview #54, BJ, 6/8/16

[81]Interview #41, BJ, 3/24/15

[82]Interview #54, BJ, 6/8/16. This source knows someone who worked on the film crew. While to protect the interviewee's confidentiality I cannot provide further details, and one must always be cautious when relying on a sole source, I consider the information highly credible.

[83]Source: Washington Post. June 9, 2016. "China's Communist Party Wants to Turn Up the Volume on Propaganda."

adherence to his preferred ideological formulations, it could also be viewed as his genuine attempt to insert people who are both loyal, and savvier in using social media than the old propaganda guard.

These anecdotes individually are not conclusive, but together raise the possibility of Xi favoring the CAC at the CPD's expense. That said, one should not overstate the case since evidence also exists that the two agencies collaborate closely. One respondent referred to the relation between the two as "two signboards, one center of authority."[84] Another key aspect is that the CAC itself is largely staffed with propaganda cadres, albeit relatively young and Internet-savvy ones; this could be viewed as Lu's and Xi's attempt to keep the CAC politically important by importing propaganda officials from Beijing municipality, while cutting out older or less savvy cadres from the CPD.[85] Although available evidence does not permit an unambiguous reading of clear intent on Xi's part to entirely exclude the CPD from Internet leadership, it clearly has lost influence.

3.6.3 Analysis: Bureaucratic Winners and Losers in the Xi Era

Figure 3.2 below summarizes the new Internet authority relations since recent reforms.[86] Where in Figure 3.1 both horizontal (among Beijing-level agencies)

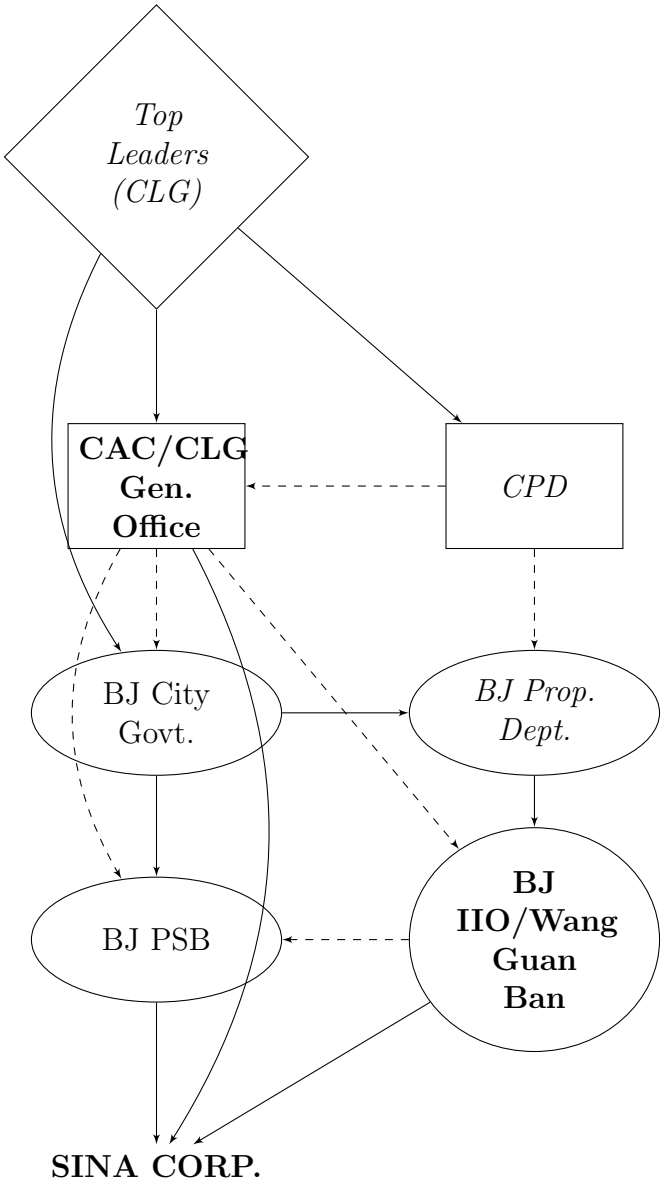
[84]Interview #37, GZ, 3/9/15.

[85]Interview #48, BJ, 4/16/15.

[86]The Beijing Municipal Propaganda Bureau maintains leadership relations over the Internet Management Office even after the reforms. Source: Xinhua. May 8, 2014. "Beijing Municipal Party Committee to Establish Internet Security and Informatization Leadership Small Group" [*beijing shiwei jiang chengli wangluo anquan he xinxihua lingdao xiaozu*]. See also the Beijing Municipal Government's 2017 Budget Report, which lists this office as receiving Beijing municipal funding: http://caizheng.beijing.gov.cn/caizheng/2801/1246561/145231/index.html.

and vertical (between Beijing agencies and the central level) *yewu guanxi* were unclear, here the CAC has clear consultative oversight (and considerable *de facto* authority to get its way) in all Internet-related matters with all other central and municipal actors, while the Beijing IIO/Internet Management Office (a.k.a. *wang guan ban*) also has consultative oversight with other Beijing agencies. The CPD, while exercising nominal authority over the entire state Internet system as a Party body, does not oversee this system *de facto*, having been superseded by the Central Leading Group and CAC.

Figure 3.2: A Clearer Hierarchy: The Chinese Social Media Censorship System Post-reform



Note: Diamond = leadership pinnacle; Rectangle = central level government agencies and Party organs; Oval = provincial (Beijing municipal) level. Solid lines = binding authority (*lingdao guanxi*); dashed lines = strong influence (*yewu guanxi*), but no binding authority. Regular text = government; Italics = Party organ. Bold text = government *and* Party organ. **Exception:** The solid lines pointing to Sina Corp. are not technically *lingdao guanxi*, since these can only exist between bureaucratic entities. But multiple interviewees indicated that Internet companies including Sina obey orders from the actors I have shown in this chart as totally binding, just as is the case for *lingdao guanxi*.

The establishment of the Central Leading Group is also consistent with the general trend of Xi using such leadership groups to circumvent bureaucratic resistance and centralize power in his own hands, ostensibly to push through anti-corruption and other difficult reforms (see Naughton 2015). These reforms have generated many potential benefits for the central state, of which two deserve note. First, the bureaucratic restructuring has nicely complemented Xi's increasing emphasis on "rule according to law" (*yifa zhiguo*), a phrase that in China could imply either actual legislative action, or rule through regulatory and administrative strengthening, provided these non-legal codes provide the Internet companies some measure of fairness and predictability in dealing with the government. That said, since in China the implementation of regulations ultimately rests on personal authority at higher levels, central-level agencies that want to be effective must enjoy the legitimacy afforded by the Party's very top leaders throwing their weight behind the reform initiative. The CAC has both, and thus is well-positioned to serve as "Internet czar" while doing so "according to law." Second, the new central-level structures complement rather than displace expertise at the provincial/municipal level; indeed, they empower this level. The CAC is able to focus on broad policy efforts and "campaigns to clean up the Web", while trusting day-to-day order-giving to the Beijing Internet Management Office/IIO, and enforcement of 'security'-relevant regulations to the Internet police.

3.7 Conclusion and Implications

Although the full extent of media and Internet system reforms under Xi Jinping has not yet fully manifested at the time of writing, two observations are clear:

1) space for online commentators is as restricted as it has ever been since social media's emergence in China; and 2) ongoing instances of swift and decisive censorship of topics the leadership deems harmful to its interests — the rapid silencing of mainland online support for Hong Kong demonstrators during the 2014 Occupy movement is one example — suggest that leaders' ability to "get what they want, and get it fast" regarding censorship implementation is stronger than ever.[87] In addition to a fierce 2013 crackdown on leading bloggers, the campaign has also affected censorship implementers themselves – both the companies, and government agents – as top officials sought to combat the phenomena of paid post deletions and what they saw as an excessive emphasis on market-driven 'hot topic' promotion at the expense of political rectitude. Employees at Sina were questioned by police, and some senior officials came under investigation.[88] Even CAC officials themselves were not immune, as some came under investigation for corruption.[89] Such efforts to clean up and reform the bureaucracy, then, have been combined with a concerted show of will by top leaders to remove unfavorable social media content: to "strike hard against rumors" (*daji yaoyan*), clean up pornography, and most recently, to "spread positive energy" – a phrase which one interviewee viewed as epitomizing Xi's clean Internet campaign.[90]

3.7.1 Alternative Explanations

This chapter has advanced three main claims that purport to explain China's

[87]Interviews: #15, BJ, 11/5/14; #16, BJ, 11/12/14.

[88]Interview #21, BJ, 11/27/14. See also People's Daily Online. 4/11/2015. "Sina faces suspension over lack of censorship."

[89]Source: Xinhua. 1/21/15. `http://www.hn.xinhuanet.com/2015-01/21/c_1114079452.htm`

[90]Interview #27, HK, 1/16/15.

success in ever-more robust censorship in the 2010s: leader beliefs about information control, a symbiotic state-company relationship, and a particular strategy of creating and elevating a new specialized agency (the Central Leading Group and CAC) to overcome resistance from entrenched existing agencies. Since I argue that the first two are necessary conditions, the implication for comparative analysis is that states lacking either are unlikely to have highly nuanced censorship programs. The Chinese case suggests that leader understandings of the value of highly responsive and flexible censorship of the sort seen during the *Under the Dome* documentary are contingent upon their pre-existing beliefs, and depend on their prior experiences with media control. Second, states lacking strong domestic Internet sectors that are beholden to the state should not be able to implement complex censorship programs.[91]

What alternative explanations could account for the apparent success of reform? One alternative is that Xi's *leadership style* and *personal beliefs* have been responsible for tightly centralized control over the Internet bureaucracy (as well as other areas). This explanation differs from the above claim about leader beliefs in that it tends to attribute events to Xi's own background and beliefs about the danger of ideological weakening rather than CCP elites' collective understanding of the information control imperative. The concern for selective censorship's necessary conditions is the possibility that as of 2012, the reform process might not have been advanced enough to be confident that censorship decisions taken on major 'hot topics' reflected strategic intent on top leaders' part and not other dynamics. The picture is further muddied by the fact that 2012 was a leadership transition year, suggesting that leader incentives to allow greater Internet openness (or al-

[91]I realize that this claim risks over-determining the Chinese case as few other authoritarian states are sizable or wealthy enough to have large domestic Internet sectors. That said, a few potential comparative examples remain: Russia, Iran, and possibly some of the larger Gulf states.

ternatively, their mere lack of attention to censorship) might explain the instances of non-censorship I observe.

While these concerns cannot be entirely refuted, it is crucial to note that since the three case studies in Chapters 4-6 focus exclusively on Sina *Weibo*, the most important organ for regulating Sina (the Beijing Internet Management Office) was already in place as of 2011 and played an active role, as did the Beijing PSB. While as identified above, numerous problems and some degree of corruption did exist within the overall censorship system, this is far from saying that top leaders, relying on Beijing-level organs for implementation, were incapable of sending out decisive orders whether to censor in response to priority online breaking events. Moreover, some major reforms, such as the separation of the SIIO from the SCIO, took place in 2011, making the reforms under Xi more of an acceleration of efforts rather than a fundamental shift in direction or a new beginning. Several interviewees supported this interpretation, seeing continuity between late Hu, and early Xi era reforms.[92] Indeed, by 2012 leaders had already begun to tighten control over the Big V without resorting to the harsher tactics they would pursue later on. Doubtless aware of this political pressure, Sina assigned personal secretaries, (*mishu*) to famous Big V in an attempt both to promote their commercial brand, and to ensure they did not cross political lines.[93] Additionally, even though major systemic reorganization did not begin until 2013, leaders pursued a series of more restrictive measures in 2011-12 even while leaving some openness for the *Weibo* 'experiment'. For example, in March-April 2012, leaders ordered Sina to turn off *Weibo's* commenting feature for three days after rumors went viral that disgraced official Bo Xilai was planning

[92]Interviews: #27, HK, 1/16/15; #29, HK, 1/22/15; #31, HK, 2/4/15; #44, BJ, 4/3/15; #49, BJ, 4/22/15. To be fair, I should note that some interviewees also added that overall reform momentum (across policy areas) was much stronger under Xi, a widely shared perception among China watchers.

[93]Interview #44, BJ, 4/3/15.

to stage a coup.[94]

A second alternative concerns the claim that leaders after 2011 had less to do with regulating the Internet *per se* than reining in those with the loudest mouthpieces online: journalists, media outlets, and prominent bloggers. In other words, while the Internet companies may have ostensibly been regulatory targets in this story, they were ultimately just intermediaries, with the real targets prominent voices who opposed Xi's program or indeed voiced anything that portrayed the Party or his reforms in a negative light. This story, however, raises the question why Xi or his associates saw the need for any institutional re-configuration or the elevation of the CAC in the first place. If control over people rather than Internet technology and the industry was really what mattered, why not just work through existing institutions like the police, and propaganda department? Of course, Xi in fact has made use of *both* existing and new institutions, with the police playing an active enforcement role in intimidating and arresting bloggers and propaganda departments creating a general ideological climate of pressure on dissenting voices. That such means have also been used, however, cannot explain the specific bureaucratic configuration we in fact observe.

A final alternative concerns the possibility that technological change, namely the rise of social media platforms as the Internet's most dynamic forum yet during the late 2000s and early 2010s, might have simply made online space a much easier regulatory target than was previously the case. In this account, while leaders ultimately took a few years (as have leaders in other countries and society overall) to grasp the power of new online forms like microblogs, once they did, these spaces proved easy to regulate because they were centrally administered by a single In-

[94]Source: The Wall Street Journal. 3/31/2012. "Sina, Tencent shut down commenting on microblogs." `http://www.wsj.com/articles/SB10001424052702303816504577314400064661814`

ternet company, and because they aggregated user-generated content into a single well-structured format, making it easy to monitor.[95] The implication here is that the complex and flexible nature of China's Internet bureaucracy should be irrelevant and thus other, less sophisticated countries should be able to replicate China's success. To be sure, the technology argument has a point in that concentration of online commentary and news into a few sites makes the 2010s Internet an easier policy area to regulate than many others. However, this argument by itself again cannot explain why leaders would see the need for extensive bureaucratic restructuring, and runs counter to numerous empirical observations: for example, the financially costly presence (in terms of salaries) of many "in-house" Internet police inside major companies like Sina. In sum, the technological change argument does play a role, but is far from accounting for the major bureaucratic re-shaping observed since 2011.

3.7.2 Broader Implications and Future Research

The findings here have implications beyond the dissertation both for the study of Chinese politics, and comparatively. First, they call into question whether fragmented authoritarianism is an appropriate framework for analyzing the Chinese Internet bureaucracy, particularly in Beijing municipality and at the central level. To be sure, the proponents of this framework have never claimed it works equally well in all policy areas, and it does not constitute a complete 'theory' of the Chinese (or any) bureaucratic system. But the fact that fragmented authoritarianism does not seem to fit well for Internet regulation is notable and admits of at least

[95]The corollary is that traditional blogs, bulletin boards and websites were *difficult* regulatory targets because of their diffuseness.

two possible explanations. First, as suggested above, the Internet has become consolidated enough that its structure is very amenable to streamlined, centralized regulation. While this explanation has some merit, a more likely possibility is that President Xi is making considerable efforts to overcome fragmentation by concentrating power at the top in his own hands (through the central leading groups) and by relying on a network of trusted, personally chosen subordinates to circumvent bureaucratic interests. This does not mean he will come anywhere close to succeeding – a large degree of fragmentation is likely endemic to bureaucracies in massive countries – but he may progress much further than his predecessors. In this sense, the Internet policy area serves as a 'cutting edge' example of just how far Xi can go in his centralization campaign. It remains to be seen whether such control is a product of Xi himself or will be transferable to whomever (eventually) succeeds him, thus allowing CCP elites to sustain robust and nuanced online information control far into the future.

Finally, the findings both help to delineate cases for comparative analysis, and direct inquiry for examples outside China. The chapter's first two claims regarding longstanding elite beliefs and the presence of a vibrant domestic Internet sector help justify why only a small subset of authoritarian states are 'comparable enough' to China with respect to online information control. However, the chapter's most important contribution is to suggest that researchers should look at what I term "traditional" versus "new" Internet regulators in these countries. "Traditional" regulators include the police and security agencies, and various propaganda authorities, while "new" ones refer to specialized agencies specifically established to head Internet regulation – various information technology ministries may also be included provided they deal with Internet content as well as infrastructure and technical standards. I argue that security, and propaganda agencies are (for very

different reasons) generally poorly equipped to implement nuanced censorship policies that allow precise and rapid variation in what is censored across specific online topics, and for using censorship as a means to complement state propaganda efforts. Comparative work can further these factors and the implications for the flexibility and robustness of states' censorship regimes.

Having provided evidence to warrant treating the Chinese state as unitary enough to implement a selective censorship logic and established leaders' motivations for such an approach, the next three chapters undertake case studies of specific online incidents using the *WeiboScope* data, to (further) evaluate both the claim of a "unitary" state, and to see whether the censorship pattern is consistent with Chapter Two's argument.

CHAPTER 4

THE BEIJING U.S. EMBASSY AIR POLLUTION DISPUTE

4.1 Introduction

Over the past two decades of China's breakneck industrial development, air pollution has become one of the most visible threats to human health, contributing to the deaths of an estimated 1.6 million people per year.[1] More recently, both pollution and government reluctance to publicize accurate air quality monitoring data have become subjects of lively debate on social media, and public attention has increased to the point where air quality has become a major political challenge to the CCP. While daily air quality (AQI) data has existed in Chinese cities since the 2000s, more recently a controversy arose over discrepancies between the U.S. Embassy in Beijing's index, which includes a measure of PM 2.5 (fine-grained particulate matter less than 2.5 micrometers in diameter, considered the most harmful to health), and the Chinese government data which only measured the larger PM 10 particulates (Chan and Yao 2008).[2] In contrast, since 2008 the U.S. Embassy has been publishing readings (including PM 2.5) taken from a monitoring station on the Embassy roof.

This chapter analyzes the government's censorship response to *Weibo* discussion about pollution in the context of a major turning point in public debate: the 2012 public confrontation between the U.S. Embassy in Beijing, and China's Ministry of Environmental Protection (MEP). Up until 2012, the MEP, which is

[1]Source: New York Times. August 13, 2015. "Study Links Polluted Air in China to 1.6 Million Deaths Per Year."

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

responsible for keeping the official government statistics, had privately urged the Embassy to stop releasing this data but had not taken further action. However, on World Environment Day (June 5), after years of private complaints about the U.S. Embassy's release of its monitoring data, MEP Vice-Minister Wu Xiaoqing finally went public, accusing the U.S. of violating China's sovereignty.[3] Wu made a number of claims. First, he assumed a nationalist posture in criticizing the U.S., using phrases such as alleged foreign "interference in other countries' internal affairs" and called on "other countries to respect our country's relevant laws and regulations." Second, Wu took a more technical approach: he pointed out that the U.S. Embassy standard for determining how much exposure to PM 2.5 was dangerous was calibrated to developed country levels and was unsuitable for China, and criticized the Embassy's reliance on a single monitoring station as "unscientific." Third, Wu clarified that he took issue not so much with the Embassy collecting the data, as that it had become widely publicized within China.

On the morning of June 6, several newspapers reported Wu's remarks from the previous day 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 PM 2.5 and the government's efforts would have taken place.[4] To make matters worse, Foreign Ministry spokesman Liu Weimin added to Wu's remarks by calling on foreign diplomats to stop issuing air quality readings, "*especially over the Internet*" [emphasis added].[5]

[3]Source: International Herald Tribune. June 6, 2012. "China tells U.S. to stop posting data on air quality; Embassy's Twitter feed on pollution is deemed improper and misleading".

[4]While the data indicate that officials generally censored such speech harshly (the overall censorship rate for June 6 was 71%), they appeared more concerned with the unfavorable contrast netizens were drawing between the Embassy and Chinese authorities, than with air pollution discussion *per se*.

[5]Source: The Vancouver Sun. June 6, 2012. "In China, pollution is not up for debate; Government orders embassies to stop issuing readings to the 'outside world'."

Weibo commentary about Wu's and Liu'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. Cui's remarks can be disaggregated into two messages. The first one attempted to deflect the public's attention away from the U.S. Embassy, saying that China could not and should not rely on foreign actors. The second message went further, arguing that the Chinese people collectively (and therefore not solely the government) were responsible for improving air pollution and implying that the root of the problem was the irresponsible vehicle use of China's upwardly mobile population.

The next day (June 13), netizen responses to Cui's two lines of argument were even more mocking than the previous episode, with netizens slamming Cui for attempting to divert blame away from what many viewed as a government cover-up. Adding fuel to the fire, prominent Big V and celebrity real-estate developer Pan Shiyi weighed in, commenting that "no one should expect the Embassy to improve air quality... first we need to know how severe the pollution is, and how much physical harm it causes... [then] remediation depends on everyone." Pan's comment was widely mocked by netizens because until that point, he had been a leading, outspoken proponent on *Weibo* for government action on pollution remediation and data transparency. Many bloggers therefore viewed him as either selling out to the government, or possibly having only issued his latest statement in response to political pressure.

The above events in June represent a turning point in *Weibo* discussion of U.S.

Embassy/MEP dispute that divides 2012 into two distinct periods or phases, where each phase contained a different government strategy for censoring discussion: a "static phase" from January until June 5, and an "adaptive phase" from June 14 through December. As the controversy ebbed and flowed across these two periods and the June 6-13 peak, 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. This pattern of censorship variation in response to the different time periods and sentiment categories is a "most likely" case of selective censorship, and provides evidence of the four-variable framework in action on an issue (air pollution) of increasing concern to *Weibo* users.

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, I show how censorship varied in response to which of these three sentiments increased on any given day, with the specific pattern differing both across the two phases, and across sentiment categories within each phase. More specifically, I find that censorship was positively associated with post surges in all three sentiment categories prior to June, but diverged after the June peak, with more hostile sentiments (toward the government) being censored more tightly thereafter and more neutral sentiments more loosely. While not excluding other variables in the framework, this pattern particularly highlights *responsiveness benefit* at work but also leaders' fear of *image harm*. The rest of this chapter elaborates and then presents results from a method for analyzing the *Weibo* data that I will apply not only here, but also in Chapters 5-6. First, however, the next section briefly discusses relevant work on the issue of environmental politics in China.

4.2 Relevant Literature

China's nascent environmental movement emerged in the 1990s after more than a decade of rapid economic growth in the Reform era. During this period, the central government prioritized economic growth and industrialization and paid little attention to environmental protection despite industrialization's massive impact on air, water and soil, not to mention the environmental degradation that had taken place during the Mao era. Environmental NGOs stepped into this breach to undertake activities that the government could or would not (Ho 2001). Somewhat uniquely among issue areas that involve organized civil society in China, these NGOs were allowed a degree of tolerance, and sometimes outright encouragement. This relative openness has led some scholars to argue that environmental NGOs (ENGOs) operate within a semi-liberalized "green public sphere" (Yang and Calhoun 2007), or are even representative of a broader "consultative authoritarianism" (Teets 2013).

In the 2000s, this green public sphere grew stronger alongside surging Internet use. While environmental campaigners attached great importance to traditional mass media, they also took to web sites, mailing lists and blogs as means to foster a "greenspeak" discourse over the decade (Ibid.) These new spaces proved effective in linking elite-level activists and volunteers together in solidarity and in helping coordinate their activities, as Yangzi Sima's ethnographic study of the prominent ENGO Global Village of Beijing (GVB) shows (Sima 2011). Yet Sima and other authors also note major weaknesses and limitations in how activists used the Internet in the pre-*Weibo* era. Elite campaigners like ENGO Friends of Nature's Liang Congjie and Green Earth Volunteer's Wang Yongchen clearly benefited from the Internet's communication properties to organize professionalized lobbying efforts

toward specific ends, like halting dam construction on the endangered Nu river in Western China (Mertha 2008). Yet they and most other environmental activists fell short in communicating their aims to and engaging the broader Chinese public.

Such a critique is not specific to ENGO Internet use, of course, but more broadly implicates conscious activist strategies to keep a low public profile in order to influence state actors while remaining within the bounds of official tolerance. Yet the "technological shock" of the Internet that has received much emphasis in this dissertation did little to change this situation in the pre-*Weibo* era. One potential exception was during acute pollution crises: for example, an explosion at a petrochemical plant in 2005 that dumped benzene into the Songhua river, which supplies water for the city of Harbin. China Central Television, which reported on the explosion immediately after it occurred, created a special web page dedicated to the incident that pooled news reporting from other sources as well as its own (Tilt and Xiao 2010). Similarly, recent "NIMBY" (Not In My Backyard) campaigns against the construction of toxic-polluting plants near residential areas – such as waste incinerator facilities (Lang and Xu 2013) – made extensive use of community online bulletin boards (BBS) to spread information about the facilities and organize opposition.

Despite such exceptions, prior to *Weibo* environmental activists and NGOs did not have either the medium, or 'household name' spokespeople to raise broad environmental and pollution awareness among the rapidly growing online population. All that changed beginning in late 2011 with Pan Shiyi, Lee Kaifu and other Big V commenting on Beijing's horrendous winter smog, and (in Pan's case) constant re-tweeting of U.S. Embassy air quality readings. In other work (Cairns and Plantan 2016b), my co-author and I undertake a more intensive qualitative study of

topic-relevant 'Big V' tweets. We find that Pan in particular stood out among the environmental activists and organizations we surveyed both for his very high follower count, and his willingness to criticize the government prior to the June turning point.[6] The importance of the Big V in sensitizing the broader public to the health and quality-of-life impact of pollution reveals *Weibo's* dual-edged nature in connecting state and society. On the one hand, *Weibo* has enabled common knowledge about China's environmental threat to form as never before, but has also prompted the state to monitor and regulate environmental discussion on microblogs much more stringently than they ever oversaw the more professional (and from leaders' perspective, less destabilizing) ENGO activities.[7] It is this mixed record of official tolerance and repression of public pollution discourse that provides the context for applying Chapter Two's theoretical framework to the 2012 dispute.

4.3 Why Air Pollution and How Was It Censored?

This chapter focuses on air pollution as topic because it represents an ideal, "most likely" case to observe selective censorship in action. This is because authorities are more likely to perceive *responsiveness benefit* in issues that a) are serious and directly affect *Weibo* users' perceived interests, but b) do not *directly* implicate top leaders' legitimacy. Environmental issues in China fit both criteria well since they occupy a privileged space in Chinese politics (Ho, 2001; Ho and

[6]Fieldwork interviews by Plantan on this point also support an interpretation of Pan's outsized role.

[7]This may have changed in the Xi era with much tighter oversight of all manner of civil society organizations, particularly foreign NGOs. But this development had no bearing on the 2012 U.S. Embassy dispute as the two time frames obviously do not overlap.

Edmonds, 2008; Yang and Calhoun, 2007; Hildebrandt and Turner, 2009), yet also arouse widespread public concern. Thus, I expect *Weibo* censorship on this issue to vary widely between tolerance and repression according to the four-variable framework's factors.

In applying the four-variable framework to the pollution issue, I 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 the air pollution topic, I term this period the state's "Static Phase".[8] 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 stages, public pressure sometimes becomes so strong that the state is prompted to reconsider its approach. Rather than doubling down on censorship, my argument is that with respect to air pollution, leaders eventually reached a turning point – an "Adaptive Phase" – where they saw the benefit of opening up selective space for tolerable criticism, while more aggressively filtering destabilizing or de-legitimizing comments. Each phase should yield different scorings of the key independent variables, and a different expected overall censorship level.

[8]I 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 I view as the state's more dynamic and adaptive response later on.

4.3.1 Explaining Censorship Variation Across the *Political*, *Physical Harm* and *Scientific* Sentiment Categories

On the independent variable side, I operationalize my 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. The method for coding the posts and their resulting sentiment categories is detailed in the next section. Here, I use those coded categories and match them to my 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, I 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. Conspicuously, *collective action risk* is absent from these predictions. This is because 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.[9] While Chinese citizens have taken to the streets to protest other environmental threats such as the construction of chemical factories (Lang and Xu, 2013; Chen, 2009), to my

[9]I 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. I relied on foreign media since these are not subject to the same in-house censorship bias as Chinese outlets. While as already mentioned, 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.

knowledge no such mobilizations have occurred in response to spikes in air pollution levels. I therefore do not think that collective action risk accounts for the censorship I observe.

I score each independent variable – my 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". I 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 4.1 and 4.2 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. I 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.[10]

Table 4.1: 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

[10]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, I do not intend to predict actual censorship levels but rather develop a categorical scheme to capture the variables' relationship with censorship's *relative* magnitude.

Table 4.2: 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, I coded *visible censorship cost* as "neutral" for the Static Phase because I 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 I 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, I expect it to be "negative" for *Political*, "neutral" for *Physical Harm* and "positive" for *Scientific*. I 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 longstanding 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, I 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 I 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 the theoretical framework. The tables predict *relative levels* of high or low censorship resulting from the independent variable scores. However, the data shows various *trends* (increases and decreases) in the sentiment categories over time. I 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.[11] These dynamic relationships should then be apparent in regression models linking daily censorship to the measures.

[11] 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.

4.3.2 Identifying Discussion of Air Pollution and Coding the Sentiment Categories

To filter out only pollution-relevant data, the 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. My co-author and I went through several stages of pre-coding exercises to determine the key categories before moving on the full coded sample.[12] 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 (Cairns and Carlson 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

[12] I jointly undertook the *Weibo* coding exercise and analysis for this chapter with a fellow graduate student, Elizabeth Plantan. A separate paper with the same data and method is intended for journal submission. However, work on the current chapter is solely my own, including literature review, theoretical logic (based on Chapter Two), results, and interpretation/conclusion.

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, the *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 the 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 empirical testing than the dates before and after them, they do enrich understanding of what led Chinese leaders to shift censorship strategy. Second, we aimed to generate year-long time series to chart the changes in the 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 and King, 2010) to estimate the proportions for the entire year. See Appendix C for more details.

4.4 Results

Before moving to regression modeling, I first present summary statistics and graphs of the estimated keyword and *ReadMe*-based proportions. Table 4.3 reports estimated mean proportions of all sentiment measures divided up into the four time periods.[13] For reference purposes, the average Air Quality Index (AQI) from the Beijing U.S. Embassy monitoring station is included.[14]

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

[14]The AQI is a composite measure of multiple pollutants, but is heavily influenced by ambient

Table 4.3: Sentiment Category Proportions, AQI, and the Censorship Rate during Peaks in Pollution Discussion

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.

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, I found that much of this news concerned local government initiatives to bring new air quality monitoring stations online. Yet however proactive these

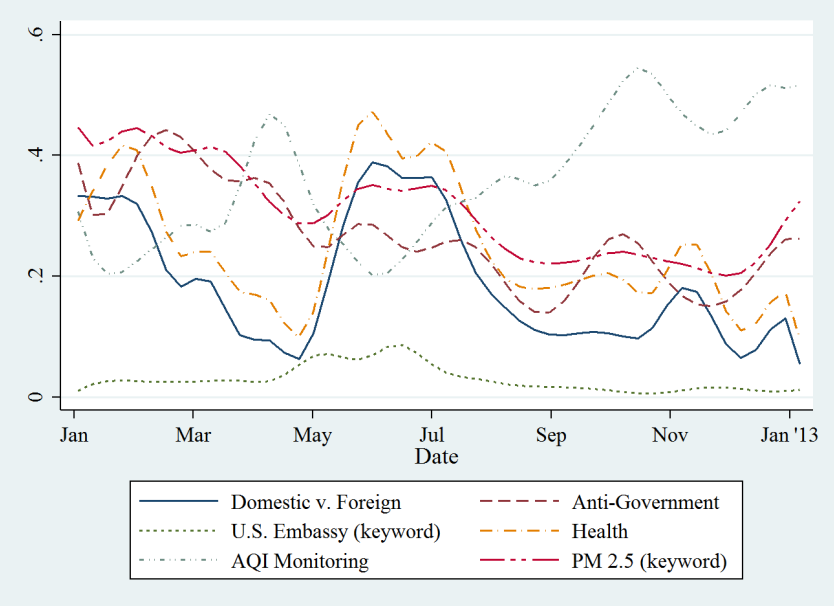
levels of PM 2.5. A 0-50 reading is considered "Good"; 51-100 "Moderate"; 101-150 "Unhealthy for Sensitive Groups"; 151-200 "Unhealthy"; 201-300 "Very Unhealthy" and 301+ "Hazardous".

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).

I next consider June 14 to December 30, which I argue represents the "Adaptive Phase" in censorship policy. The *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, the *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, I examine year-long graphs of the hand-coded and keyword proportion estimates in Figure 4.1.

Figure 4.1: Sentiment Category Proportions Across 2012: Air Pollution Dispute



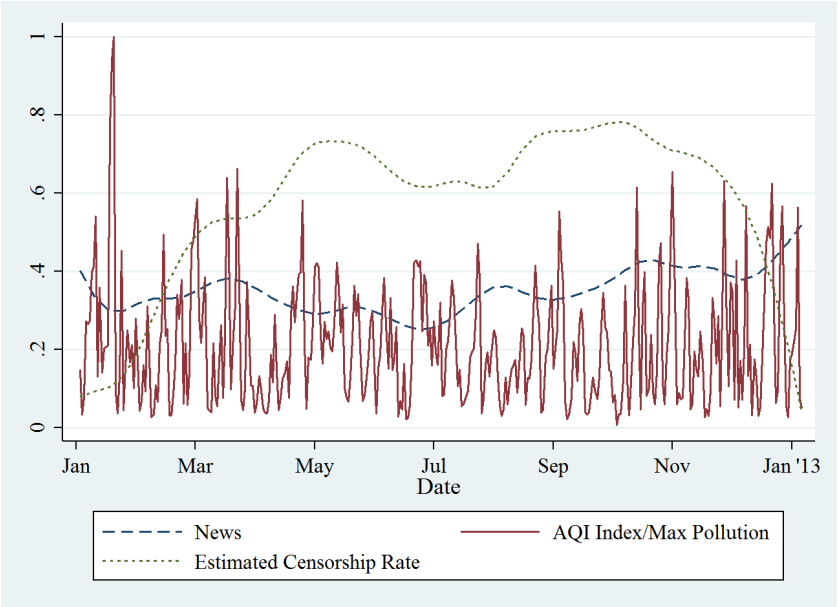
Looking at the graph, I notice a surprising finding: a strong correlation between Domestic vis-à-vis Foreign and Health, which I had expected to diverge since they were supposed to represent two different sentiment categories. The reason for such a correlation was not immediately evident, but became clearer upon a qualitative reading of the data. While during coding my co-author and I 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 I explore further in the 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.[15]

Finally, I examine the control series, and also include the dependent variable of the censorship rate in Figure 4.2:[16]

Figure 4.2: Proportion of News, of AQI/Max AQI, and of Posts Censored Across 2012: Air Pollution Dispute



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

[16]This graph shows an *estimate* of the censorship rate subject to some key assumptions (see Appendix B). While I am 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 my 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.

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) 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.[17] Yet censorship's overall trend is not inconsistent with my 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 the 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 4.2 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, I 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.[18] Using this ratio, I 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 because air pollution was visibly bad, which is something that would be difficult to cover up through censor-

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

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

ship. 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 I model the relationship between these proportions and the censorship rate.

4.4.1 Modeling the Sentiment Categories' Relation to Censorship

In this section, I consider the statistical relationships between the sentiment measures and the censorship rate. To do this, I compare regression models for January 2 - June 5 with those for June 14 - December 30, the periods before and after the June peak that I argue shifted censorship policy. While I do not make specific predictions for each coefficient, in general I expect the directions of significant effects to resonate with Tables 4.1 and 4.2: *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 the measures consist of time series, I 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 the dependent variable is a proportion, while

OLS and other models assume the dependent variable can take on any real number. To address these problems, I use Generalized Linear Model (GLM) regression and assume that the censorship rate has a binomial distribution and that the model takes a logistic form. I then deal with autocorrelation by employing Newey-West standard errors. Newey-West models require specifying the model's maximum lag order, for which I rely on the Akaike Information Criterion (AIC). Based on the AIC results, I chose a lag order of four.

I incorporated this information into my model in two ways. First, I 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, I set the error term maximum lag at 4 – this should control for any residual interdependence among the series not accounted for by the included variables. Taking first the Static Phase, Table 4.4 shows average marginal effects.

Table 4.4: Sentiment Category and Keywords' Relation to the Censorship Rate (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

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

The table presents four model specifications and displays lags zero and one.[19] Model I consists only of the key independent measures for *Political* and *Scientific*; the *Physical Harm* measures are absent from the baseline model because I found that with one exception, none of them were significantly related to the censorship rate. Nonetheless, I kept them in the analysis in Model IV, as they might still be (weakly) correlated with my dependent and independent variables. Model II adds News, and Model III further adds the AQI Index. Looking at the results, I immediately note that lag one of the censorship rate is positive, significant and large. Given my prior understanding of censorship as typically reactive with some lag to sudden bursts of online controversy, I was 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 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* – the one exception to overall null findings for *Physical Harm* – alongside the lack of significance for U.S. Embassy in Model IV, I 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

[19] Although the actual regressions were run with lags two through four also included, and the coefficients in Tables 4.4 and 4.5 reflect this influence, I omit reporting these results for brevity's sake and because I are interested only in more recent lags' effect on censorship.

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 my 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, my 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 the 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 messages more explicitly critical of Chinese leadership.

The other key results for Table 4.4 concern AQI Monitoring and PM 2.5, which are consistently negative and significant across lag one. Across both phases, I found that PM 2.5 was a consistently significant predictor; indeed, it represents the best measure available with respect to capturing *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 4.5: Sentiment Category and Keywords' Relation to the Censorship Rate (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

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

I now compare the results for the Static Phase to those for the Adaptive Phase in Table 4.5. As with Table 4.4, 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 I 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 4.4, only this time at lag zero instead of one. The consistency of this measure across both time periods is unsurprising since I 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. I 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 4.5 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 I cannot claim with certainty that increased PM 2.5-related speech caused reduced censorship, I can assert that at the very least, surges in PM 2.5 discussion from June 14 onward did not correlate with *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 data to show the government's responsiveness to public demands. These results support my claim in Table 4.5 of a divergence between how the government censored *Scientific* versus *Political* sentiments after June 13. Finally, in contrast to Table 4.4, *all Physical Harm* measures are insignificant. I

interpret this to mean that these measures did not overlap with *Political* as they did earlier.

4.5 Conclusion: A Clear Shift in Category-specific Censorship Across Time Periods

In conclusion, I find that my theoretical predictions from Tables 4.1-4.2 are generally well illustrated by the data. I 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 the PM 2.5 measure. Concerning the *Political* category, while I 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 the main results were consistent with my predictions, there were a few surprises. First, I was 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 I did not specifically foresee this result and so cannot consider it validation of my 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 chapter's scope, this tentative analysis does suggest 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, as already mentioned in Chapter Two, this finding requires a few caveats. I am *not* claiming that leaders are able to micro-manage the bureaucrats and company censors that actually oversee post deletions and keyword blocking. Nor do I 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.

With these limitations 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 would be able to issue fine-grained orders that selectively filtered 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. 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.

4.5.1 Broader Implications and Future Research

Overall, this chapter illustrates how online censorship in China can vary based on both the 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 the 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.

While Chapter Four has shown an "easy case" for where clear *responsiveness benefit* (and therefore, a positive *credibility payoff*) can matter, most political issues on Chinese social media are not as openly discussed as air pollution. The next chapter considers the downfall of Politburo official Bo Xilai, a case that implicated Chinese leaders and the Party-system at the highest levels and where *credibility payoff* was often negative. Nonetheless, the amount of censorship still varied over the scandal's several-month timeframe, a development I argue was due to the scandal's high *visible censorship cost* at various points.

CHAPTER 5

THE BO XILAI SCANDAL

5.1 Introduction

Chinese Communist Party leaders' decision in April, 2012 to investigate top official and Politburo member Bo Xilai on charges of corruption and complicity in the murder of a British businessman, Neil Haywood, sent shockwaves through both official and social media channels right before a crucial leadership transition during the 18th Party Congress that year. In the scandal's early weeks before Party leaders settled on an official line, news outlets engaged in a flurry of reporting far more diverse than during previous instances of high-ranking official malfeasance. Allegations of Bo's misdeeds ranged from bribery, to illicit sexual activity, to supposedly plotting a central-level power grab, not to mention involvement in Haywood's murder. Yet until the official announcement that Bo would be removed from his post as Chongqing Party Secretary on March 15, and to some extent even until his removal from the Politburo on April 10, all of the above narratives competed in a cacophonous online sphere, particularly on microblogs like Sina *Weibo*, at the time China's preeminent venue for viral discussion of current events.

While raucous, disjointed and ultimately full of mis- and false information, these discussions carried high stakes for the CCP in shaping how the online public would view the deeper meaning of Bo's downfall. At least three explanations competed to account for Bo's removal from his posts and ultimately from the Communist Party. First, the main explanation promulgated by top leaders was simply that Bo was a criminal who had abused his power as Party secretary to

engage in various misdeeds, culminating with his involvement (along with his wife Gu Kailai) in arranging Heywood's murder. Branding Bo as a rotten official served top leaders' purpose of deflecting attention away from other, more political motives they might have for opposing him.

A second, more politically charged explanation had to do with Bo's alignment with China's "New Left" (sometimes also referred to as "neo-Maoists") while running Chongqing. Bo had championed public displays of nostalgia for the Mao era and a return to overt leftist politics in public life, such as encouraging the singing of "Red" songs in schools and workplaces to revive revolution-era communist ideals, a practice largely rejected by Chinese leaders ever since Mao's death. This view of Bo's defeat held that leaders removed him to prevent the spread beyond Chongqing of such discredited practices.

Finally, the third and most potentially damaging theory (from leaders' perspective) was that Bo was removed because he had broken an unwritten but powerful norm of elite-level Chinese politics in the Reform era: not to jockey overtly for top posts, especially membership on the Politburo Standing Committee, the center of power in China. Variants of this theory ranged from the fairly innocuous – Bo's unconventional, attention-grabbing style and populist reforms in Chongqing such as providing public housing – up to serious threats to Party unity, especially the revelation that Bo had ordered his police chief Wang Lijun to wiretap senior leaders' communications, including President Hu Jintao's.

As these theories and related attempts by ordinary citizens and media professionals alike to establish the "facts" percolated through social media, different and often opposing viewpoints emerged in netizen comments. Especially in the scandal's first few weeks, many netizens rose to Bo's defense, viewing allegations

against him as central leaders' attempt to eliminate a political rival and to strike a blow to neo-leftism more broadly. Other bloggers, however, acknowledged Bo's wrongdoing but differed as to how it reflected on CCP leaders and the political system. While some commentators accepted the idea that Bo was just a "bad apple", others went further to express doubt or cynicism about central leaders' true motives, or even declared the Party as a whole irremediably illegitimate, with Bo just representing the "tip of the iceberg" of a systemic problem.

As in Chapter Four, this chapter analyzes leaders' censorship response to the shifting costs and benefits captured by the four-variable framework during a major incident. I find that during the Bo scandal, the censors actively deleted comment threads that used the scandal to broadly question the Party-state's legitimacy to govern, or even expressed skepticism or cynicism toward central leaders' true motives. On the other hand, especially early on a surprising amount of discussion that aimed merely to find out the "truth" or facts of the case went uncensored despite its focus on such a sensitive topic. Moreover, even voiced support for Bo – risky for central leaders because of its association with the Maoist New Left – was censored more heavily at some points than others. The pattern of censorship in this case, which was strongly shaped by highly visible news events, especially highlights *visible censorship cost* in action, although not to the exclusion of *credibility payoff*.

The following sections further develop the linkage between theory and the Bo case, using the theoretical framework to make specific predictions concerning the rate of post deletions in each of three "phases" of the scandal – I) the immediate aftermath of Wang Lijun's flight to the U.S. consulate in Chengdu; II) the time period surrounding initial action by central Party leaders to remove Bo as Chongqing Secretary and then to dismiss him from the Politburo, and III) Bo's

expulsion from the party months later in September, 2012 and its aftermath. Each phase coincided with a surge in news and *Weibo* commentary but showed different patterns of daily censorship. Using methods nearly identical to Chapter Four's, I identify the major viewpoints toward Bo and the Party that prevailed at different moments during the scandal, and manually code sample posts accordingly. I then use *ReadMe* to apply this categorization scheme to a much larger body of posts taken from each of the three phases, using the hand-labeled data as "training" input. This exercise yields estimates of the breakdown of category proportions on any given day, which I then interact with the censorship rate in time-series analysis, providing a means to evaluate the theoretical framework's validity. Before beginning this procedure, though, the next section briefly mentions prior work on elite politics, corruption and leadership transitions as well as specific work on the Bo scandal.

5.2 Relevant Literature

Factionalism at the pinnacle of the CCP has received much scholarly attention as a supposed threat to intra-elite unity and a barrier to strong centralized rule (Miller 2015; Wang 2006). It has been a major target of Xi Jinping's efforts to emerge as China's unquestioned "core" leader, a status enjoyed by neither Hu nor his predecessor Jiang Zemin. Whether or not we attribute patterns of intra-elite competition for Politburo and Standing Committee seats leading up to Party congresses as a product of organized factions, however, the question of how to maintain elite unity and to discipline top-level Party members is a crucial one to understand what CCP leaders view as one of the greatest threats to continued

Party rule. In this context, both the lenses of factionalism and of elite-level Party discipline are potentially useful in sizing up why leaders viewed the Bo scandal as a major impediment to a successful 18th Party Congress and transition that necessitated a decisive response.

First, using the lens of factional politics, Bo was threatening because of his and his family's longstanding allegiances. His father Bo Yibo had played a major role in promoting Jiang Zemin's rise to CCP General Secretary and the Bo family had received Jiang's patronage, positioning Bo Xilai in the early 2000s as a long-term candidate to join the Politburo and to vie for a Standing Committee seat (in competition with Xi Jinping and Li Keqiang, who later won out as China's number-one and number-two officials). Bo also enjoyed the support of Standing Committee member and powerful security portfolio-holder Zhou Yongkang, who reportedly cast a lone dissenting vote when the Committee later met in March 2012 to decide Bo's fate.[1] During the scandal, some authors (Yuen 2014; Fewsmith 2012) noted the importance of such patronage networks in Chinese politics and cited Bo's membership in a losing network (that of Jiang Zemin) as a potential contributor to his downfall – Bo was simply on the wrong side as Xi ascended to power, and like other Jiang associates needed to be dealt with as a rival (including Zhou, who like Bo was later charged with violating Party discipline.

Although compelling, the factionalist account overlooks a major trend in Chinese elite politics in the reform era: a move away from Mao-era "winner take all" politics in which losing pretenders to the throne were simply detained or eliminated without at least some minimal basis in reality. Instead, Central Commission for Discipline Inspection (CCDI) officials and top leaders were careful to enumerate

[1]Source: The New York Times. 3/29/12. "China's Hierarchy Strives to Regain Unity After Chongqing Leader's Ouster."

Bo's alleged crimes, including the explosive report that he had wiretapped President Hu and other leaders, and his involvement in Heywood's murder.[2]. More broadly, they criticized Bo's "Chongqing model" of combining Mao-style propaganda with populist economic policies. In short, the litany of accusations about Bo's illegal activities, policies and leadership style that played out in news outlets during the scandal suggest a second, individual-level explanation for his downfall – his violation of numerous (often unwritten) governing rules and norms expected of top Chinese leaders, who are normally conservative and cautious in style and at least publicly deferential to superiors.

Along these lines, Gueorguiev and Schuler (2016) find that would-be top leaders in China and Vietnam who become exceptionally well-known are less likely to be promoted even if they enjoy strong patronage and are perceived as competent, a finding the authors argue is due to the threat to one-party rule of candidates with large personal followings. Broadhurst and Wang (2014) echo this in their analysis of Bo's tenure in Chongqing and campaigning for a spot on the Standing Committee, noting his ruthless and ambitious maneuvers to raise his profile. And as Yuen (2014) notes in comparing Bo's removal with Zhou Yongkang's, both individuals' downfall as well as that of other "tigers" can be seen as part of a broader struggle by Xi to purge the CCP of corruption and improve discipline, a broad rubric under which he means not only criminal behavior, but any sort of individualistic style in governing or disloyalty to the center.

For purposes of analyzing state censorship of social media, which lens (factional, or Bo's style and actions/Party discipline) is the most "correct" in explaining his downfall is not important: the above analysis serves merely to illustrate that top

[2]Source: The New York Times. 4/26/12. "Fall of Chinese Official Is Tied to Wiretapping Of His Fellow Leaders."

leaders under soon-to-be-President Xi Jinping had ample reason to carefully control online commentary that might raise questions about Party unity or Bo's relation to other elites. Thus, the scandal represents a "hard case" for selective rather than blanket censorship, since such questions began to spread online right after news of Wang Lijun's flight to the U.S. Consulate broke. Supporting this, numerous empirical findings document rigid censorship of topics that involve top leaders or Party factions. In their landmark study, King, Pan and Roberts (2013) find evidence of pervasive and rapid post deletion in the weeks immediately following Wang's flight. Similarly, Fu, Chan and Chau (2013) find in their *Weibo* data sample that the keywords "corruption" and "Wang Lijun" were blocked in 2012.

Why, then, would we expect to find selectively lighter censorship at any point, despite Chapter Two's argument? I argue that the answer lies in leaders' intent to persuade the online public of their seriousness in combating corruption at the highest levels, and in their knowledge that censoring the incident early on would just fuel negative speculation about Bo, Wang, and broader CCP corruption. Temporarily allowing limited social media discussion during the scandal's early days could facilitate both goals. The next section develops theoretical predictions for each of the scandal's three main phases, each of which yields hypotheses about the relation between daily censorship and different sentiment categories.

5.3 Why the Bo Scandal and How Was It Censored?

In comparison with Chapters Four (air pollution) and Six (nationalist protest), the issue of elite-level corruption and investigations is an unlikely one for selective censorship given its direct implications for the Party's and top leaders' own image

of strength and unity. Thus, overall we should expect to observe high censorship throughout the scandal, which in turn should be correlated with a broad range of different discussion topics. However, since the state is still sensitive to the shifting costs and benefits of *responsiveness benefit*, *image harm*, *collective action risk* and *visible censorship cost*, we should still expect to observe disaggregate within-case variation across time periods and topics/categories.

In the next section I discuss how I derived and coded the sentiment categories, but deal first with the temporal aspect here by breaking down the scandal's three phases. Phase I (February 8 - March 8) involved Wang's trip to the U.S. consulate and a few weeks thereafter, a time period in which Wang's connection to Bo, the extent of Bo's deeds and what action the higher-ups in Beijing were planning to take were all unanswered questions. A general prediction for the state's censorship response is difficult to make for this period because it is difficult to determine at what point top officials adopted a coherent and definite Internet management strategy for this issue, possibly because of initial divisions among leaders over how to handle the case. However, because in other cases officials have historically issued censorship orders to media within days of even highly complex political scandals, attempting such predictions is not outside the realm of possibility, and I will endeavor to do so in the next section.

Second, Phase II (March 9 - April 17) represented the pivotal moment in Bo's downfall, and involved a series of turning points. First, on March 7, the Politburo Standing Committee adopted a decision to dismiss Bo as Chongqing Party Secretary. Then a week later on March 14, Premier Wen Jiabao criticized Bo during his annual press conference, rebuking Bo's attempts to revive "red culture" in Chongqing. Finally, on April 10, Bo was suspended from the Politburo and

Central Committee and officially put under "investigation for serious disciplinary violations." Additionally, Bo's wife Gu Kailai was named as a suspect in the death of Neil Haywood. Compared with Phase I, reading fluctuations in censorship as a matter of state intent in Phase II is relatively more tenable because from March 7 onward, leaders had agreed how to deal with Bo and likely had an idea about what sort of public discussion to permit, although they could not have foreseen the exact degree and nature of public reaction to various news events.

Finally, Phase III (September 17 - December 30) differs substantially from the first two phases in that it occurred months after Bo's initial downfall, and well after his removal from power was certain. On September 28, the Politburo adopted a decision to expel him from the Party.[3] Among the three time periods, I expect Phase III to show the strictest overall censorship, and even to observe the absence or near-absence of more sensitive sentiment categories. By September, Party leaders had forged consensus not only to oust Bo from all public posts and the Party, but to publicly repudiate him, list his alleged crimes and begin judicial proceedings against him. A major motivation for these decisive actions was the 18th Party Congress, scheduled to begin in November, where Xi Jinping would officially become China's number one leader. Leaders doubtless viewed tying up the non-judicial stage of the Bo affair (by expelling him from the Party) as necessary to avoid further distraction during the Congress.

[3]I set Phase III's starting date to September 17 instead of September 28 because discussion about Bo's ultimate fate was already beginning to surge by the former date after a months-long lull, marking a new phase in public attention to the issue.

5.3.1 Explaining Censorship Variation Across the *Supporters*, *Questioners* and *Critics* Sentiment Categories

Just as in Chapter Four, I operationalize Chapter Two's four-variable 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. My coding method for the sentiment categories is detailed in the next section.[4] As in Chapter Four, I 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.

I coded three sentiment categories in this paper corresponding to three different groups of individuals with respect to the Bo Scandal: *Supporters* (those who praised Bo's policies and achievements or defended him against what they viewed as a witch hunt or purge, often but not necessarily from a leftist or neo-Maoist perspective); *Questioners* (those who expressed curiosity about finding the "truth" or analyzing the scandal's events); and *Critics* (those who went beyond narrowly targeted criticism of Bo to raise doubt or skepticism about the rectitude of top leaders' intentions in pursuing him, or even used the Bo case as an opportunity

[4] As with Chapter Four, *collective action risk* is absent from the analysis, being essentially constant and low throughout the Bo scandal. However, the reason why differs from Chapter Four, which focused on the issue of air pollution. While pollution-related protests might theoretically be tolerated, air pollution historically has not shown potential to motivate people into the streets. In contrast, during the Bo scandal there well could have been no shortage of individuals willing to take to the streets (supporting or opposing Bo), but given the issue's extreme sensitivity, most citizens would understand that any incipient street protests on this topic would be immediately and ruthlessly suppressed; the state's credible threat of using force in this case would keep collective action risk low.

to challenge the Party or system more broadly). Then, in Tables 5.1-5.3 below I score each category on a scale from -2 to +2, with -2 as "Very Negative", 0 as "Neutral" and +2 as "Very Positive". I 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 5.1-5.3 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. I 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.[5]

Table 5.1: Predicted Censorship by Sentiment Category (Phase I: February 8 - March 8)

Category	Cred. Payoff $(-2x)$	Vis. Cens. Cost $(-x)$	Pred. Censorship
Supporters	Neutral (0)	Positive (+1)	-1
Questioners	Positive (+1)	Positive (+1)	-3
Critics	Negative (-1)	Positive (+1)	+1

Table 5.2: Predicted Censorship by Sentiment Category (Phase II: March 9 - April 17)

Category	Cred. Payoff $(-2x)$	Vis. Cens. Cost $(-x)$	Pred. Censorship
Supporters	Negative (-1)	Very Positive (+2)	0
Questioners	Neutral (0)	Very Positive (+2)	-2
Critics	Very Negative (-2)	Very Positive (+2)	+2

[5]As in Chapter Four, I do not intend to predict actual censorship levels but rather develop a categorical scheme to capture the variables' relationship with censorship's *relative* magnitude.

Table 5.3: Predicted Censorship by Sentiment Category (Phase III: Sept. 17 - Dec. 30)

Category	Cred. Payoff ($-2x$)	Vis. Cens. Cost ($-x$)	Pred. Censorship
Supporters	Very Negative (-2)	Neutral (0)	$+4$
Questioners	Negative (-1)	Neutral (0)	$+2$
Critics	Very Negative (-2)	Neutral (0)	$+4$

Beginning with Phase I, I coded *visible censorship cost* as Positive because Wang Lijun's dramatic flight and the almost immediate media coverage it received made the incident highly visible to politically-inclined Chinese Internet users, and made subsequent discussion of Bo's connection to Wang more difficult to suppress regardless of specific sentiment. For *credibility payoff*, I coded *Supporters* as Neutral due to what likely was the Party's ambivalence toward Bo's leftist supporters: silence them too quickly and leftists might view their leaders in Beijing as betraying the Party's Maoist heritage by betraying Bo; let supporters talk indefinitely and they might shift public opinion in Bo's favor and against leader efforts to remove him. I coded *Questioners* as positive because leaders would be *relatively* likely to relax censorship for this category (compared with the others) especially early on in the unfolding scandal, under the logic that doing so would implicitly signal to netizens that their speculations about Bo's guilt were valid and reflected top leaders' own view. Finally, I coded *Critics* as Negative because even during this relatively open phase, leaders would be on the lookout for the possibility that bloggers might already be using the Wang/Bo affair to raise broader questions about the Party overall or the political system's integrity.

For Phase II, I coded *visible censorship cost* as Very Positive because the scandal had become major national news, with both mainstream and fringe outlets reporting on each new development. Top leaders' decision to remove Bo from his

post and later to investigate Gu Kailai could not be done quietly, despite what they might have wished. For *credibility payoff*, I coded *Supporters* as Negative – by this point, leaders had reached agreement to move against Bo and thus online supporters posed an increased risk of countering their message. *Questioners* were coded as Neutral: their speech did not serve the purpose, as earlier, of convincing citizens where leaders stood, but my expectation here was that officials probably were prepared to tolerate some further blogger curiosity about the case stemming from the March and April announcements. Finally, *Critics* were coded as Very Negative since by this point, leaders likely perceived the danger of such "tip of the iceberg" arguments about Bo to have escalated.

Last, for Phase III I coded *visible censorship cost* as Neutral: on the one hand, news of Bo's expulsion from the Party had to be made public and would attract some attention, but on the other, the completion of Bo's downfall (losing Party membership, and then potentially facing criminal charges) had likely been treated as inevitable by most online citizens for months.[6] As such, news that he had finally been expelled from the Party, while noteworthy, would not have been surprising to many. For all three sentiment categories, I then coded *credibility payoff* relatively negatively, with *Supporters* and *Critics* Very Negative and *Questioners* Negative; the only reason for coding the former two sentiments relatively more negatively was their more politicized nature compared with *Questioners*. In general, however, I expect all three sentiments to be associated with higher censorship during this phase.

As in Chapter Four, the next step is then to link predictions of censorship levels

[6]In China, the public announcement that a high-ranking official is under investigation for "serious disciplinary violations" typically leads to that individual's loss of Party membership, followed by criminal procedures of some kind. In Bo's case, the assumption that all these steps would take place would have been especially widespread given Bo's rank and notoriety.

to these dynamic trends by treating the latter as observable implications of the former. These dynamic relationships should then be apparent in regression models linking daily censorship to the above categories. First, though, I address how I isolated posts relevant to the Bo scandal from the WeiboScope data, and how I defined and coded the sentiment categories.

5.3.2 Identifying Discussion of the Scandal and Coding the Sentiment Categories

To filter out only scandal-relevant data, my sample consisted only of posts containing one or more of the following keywords (in Chinese characters): "Bo Xilai", "Wang Lijun", "Chongqing Sick Person" (*chongqing bingren*, a euphemism for Wang Lijun, based on a documentary title), "very open news policy" (*duome kaifang de xinwen zhengce*, a phrase I found to identify topic-relevant content related to netizens' observation of relatively low censorship in the initial weeks after Wang's flight), "Secretary Bo" (his official Party title as Chongqing leader), "Discipline Inspection Commission" (referring to the Party disciplinary body that brought an investigation against Bo), "Central Discipline Inspection Commission" or "CDIC" (abbreviated *zhong ji wei*), "Gu Kailai", "Bo Gu" (the two surname characters for Bo and Gu, respectively, which often appeared in news reports as a single unit),[7] "serious disciplinary violations" (*yanzhong weiji*), "the wife of

[7]Mysteriously, state media reports about Gu Kailai's involvement in Heywood's death and Bo's connection referred to her surname as "Bo-Gu", a manner in which she had never been publicly addressed before. In mainland China after 1949, women have generally used their own family name rather than their husband's in public; before this, it was customary in some instances to add the husband's surname similar to the Western tradition. The Communist takeover in 1949 nearly wiped out this practice, a fact remarked upon by many netizens who found it bizarre that Gu would be referred to with her husband's surname added. Many speculated that this was the news media's attempt to intertwine the two individuals' misdeeds in public consciousness.

Comrade Bo" (*bo xilai tongzhi qizi*), "Comrade Bo", "Heywood" (referring to Neil Heywood), "expulsion from public office" (*kaichu gongzhi chufen*), and "expulsion from the Party" (*kaichu dangji*). This left 68,885 total posts across Phases I-III in 2012. I went through several stages of pre-coding exercises to determine the key categories before moving on the full coded sample. Appendix A details the procedure.

After multiple reads of post samples, I defined each sentiment category as follows. First, *Bo Supporters* included two groups of comments. One group backed Bo as an exemplar of Maoist ideals, using his position in Chongqing to fight for ordinary people's interests. The second group, while not necessarily Maoist or leftist, believed that Bo was an honest official who made sincere efforts to fight corruption. Although these two groups represented different ideological backgrounds and were originally coded separately, I ended up subsuming them into the *Supporters* category; I judged that state censors would not be as concerned about commenters' exact reason for support as about the mere fact that they were doing so.

Second, *Questioners* were a broad group that included individuals with all manner of queries about the nature of the scandal. While some commenters merely wanted to know what was going on, others took a more aggressive approach and insinuated that Bo was guilty, and demanded further information. Of course, this category evolved between Phases I-III as additional information about Bo's and Gu's misdeeds became public. At each stage, however, comments in this category were characterized by posters' desire to learn the facts, and in many cases to analyze them in depth. I excluded posts where I felt that commenters' factual questioning or analysis went beyond a focus on Bo, and spilled over into broader criticism or cynicism about the political system or its leaders.

Third, *Critics* included a range of different comments. Posts that might have otherwise been counted as *Questioners* but whose comments I judged to have expressed skepticism or cynicism ended up in this category. More blatant were posts that lamented leaders' takedown of Bo as just business as usual in the hard-scrabble world of Chinese politics, hinting that Bo was merely on the losing end of a factional struggle in which allegations of wrongdoing or corruption were merely a weapon, or even worse, that top leaders had gone after Bo to deflect attention from their own corrupt tendencies and poor governance practices. Even more provocative were posts that outright attacked the Party-state as rotten to the core, or that issued clear calls for systemic reforms like judicial independence or constitutionalism. I originally coded the latter two instances as separate categories, and subsumed them under the broader rubric of *Critics* as a practical matter only because these other categories had too few posts to reliably estimate their proportions, or for use in computer-assisted analysis.

Finally, I coded two "residual" categories that did not fall under one of the above three sentiments. *News* consisted of all instances in which a blogger (or news organization with a *Weibo* account) merely reposted what was clearly professional journalist or opinion content. This included both state and commercial media, as well as online news portals like Netease and Tencent. A major exception were those posts that contained a news "re-tweet" but also included some original content that fit into one of the other three categories, in which case I assigned the post to that category (if the post had news plus a comment that did not obviously belong to another category, I left it as part of *News*).

Second, I then coded all other posts not assigned to one of the other four categories as *Other*.[8] The most common potentially theoretically relevant sentiment

[8]While originally I had separated out posts deemed to be topically irrelevant from relevant

that ended up on *Other* was what I termed "mainstream" criticism of Bo, i.e. posts focusing on Bo's bad behavior but without any hint of doubt, skepticism, or a broader anti-system critique. I had initially expected such posts to be frequent and during initial coding, treated them as a distinct category. However, to my surprise as coding progressed, I found such posts to be infrequent enough so as to make a *Mainstream* category un-analyzable statistically. Realizing this, I went back to the posts I had coded as *Mainstream* and re-assigned all of them to other categories as appropriate, with the majority ending up in *Other*.

This exercise had two goals. First, I wished to estimate the proportion of posts during the peak moments within each broader phase (I-III). Doing so provided a summary statistic into which sentiment(s) predominated during key junctures throughout the year, and each corresponded to a news event. Phase I had its peak around February 8-12. Accordingly, I manually coded a random sample of 250 posts drawn evenly from across these five days into the sentiment categories. Second, Phase II had two separate peaks: March 14-16, and April 11. Even though I lumped these two periods into a single phase as the two news events were related (Bo's removal as Chongqing Party chief, and the continuation of his downfall via the accusations against Gu), it was important to code each peak separately in order to characterize the phase overall. I thus drew and coded two separate samples of 250 posts each, one from March 14-16 and the other from April 11. Last, Phase III had a single peak on September 28-29, and as with the other peaks I drew a total of 250 posts from these two dates. Thus, altogether my sample consisted of 1000 posts taken from four different date ranges, a size I judged sufficient for reasonably precise directly estimated category proportions for each peak, and for use in *ReadMe*.

posts that expressed some sentiment not captured by the four categories, I ultimately chose to combine these two categorizations due to the infrequency of irrelevant posts.

5.4 Results

Before moving to regression modeling, I first present graphs of the estimated censored posts and *ReadMe*-based proportions. Table 5.4 reports estimated mean proportions of all sentiment measures divided up into the above four peak dates.

Table 5.4: Sentiment Category Proportions and the Censorship Rate during Peaks in Scandal Discussion

Measure	2/8-2/12	3/14-3/16	4/11	9/28-9/29
Supporters	.10	.10	.05	.03
Questioners	.21	.09	.26	.09
Critics	.19	.15	.20	.16
News	.25	.52	.34	.56
Other	.24	.14	.15	.16
Censorship Rate	.23	.52	.40	.69
Daily Avg. Posts	3251	2327	2385	1175

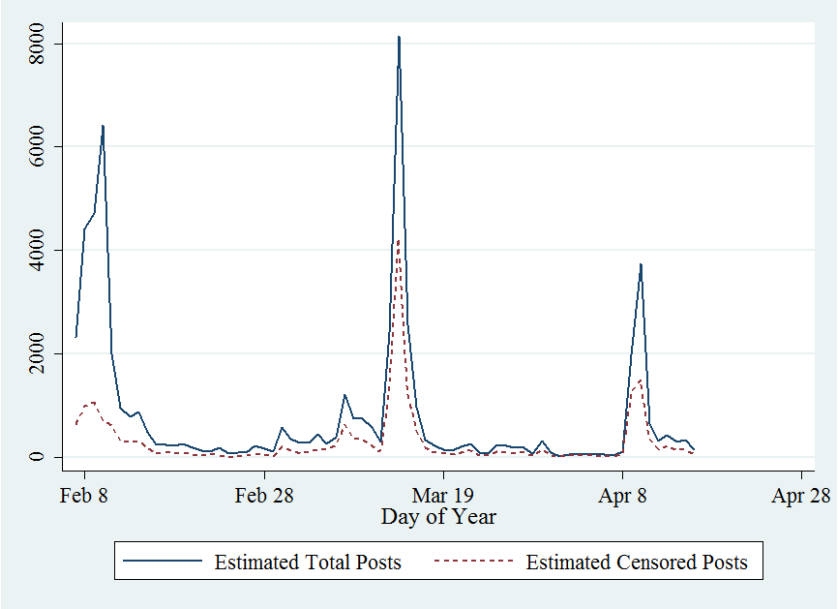
Several trends stand out. The proportion for *Supporters* declines across time periods. *Questioners* varies across peaks, a result I suspect is due to this category being the most news-driven (the first peak (Wang's flight) and the third one (the announcement of investigation into Bo and into Gu Kailai) were in many ways more scandalous and shocking than the other two). A similar trend obtains with *Critics*. *News*, on the other hand, prevails during the less-volatile news events of March 14-16 and September 28-29; of note, the presence of more news coverage and less independent commentary is positively related with the censorship rate.

These summary statistics are able to provide a rough topical breakdown of online discussion during different periods. For example, a *Critics* proportion of 19-20%, while not high, is fairly robust for similar political incidents on the Chinese Web, especially given the sensitivity of the Bo case. And Supporters dwindles to

near zero by the end of the year, as might be expected given the trajectory of Bo's downfall. However, to get more leverage on these trends, time-series graphs are useful.

Figure 5.1 below shows estimated total and censored posts for Phases I-II.[9]

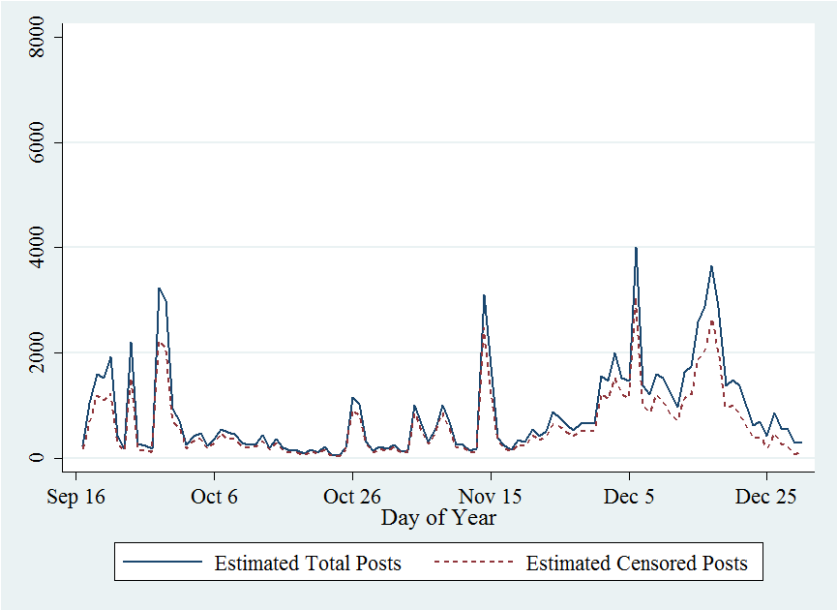
Figure 5.1: Total and Censored Post Volume: Bo Scandal (Feb 8 - Apr 17)



The graph identifies the middle of Phase II (March 14-16) as the year's highest peak, but with relatively high censorship. In contrast, the initial surge around February 8-12 (Phase I) had fewer posts but very low censorship by the standards of *Weibo* incidents. Finally, the peak on April 11 was lower, briefer, and in the middle in terms of censorship. Overall, the graph supports my initial prediction of a low censorship rate in Phase I and a moderately high one in Phase II. I now turn to Figure 5.2, which estimates total and censored posts for Phase III:

[9]Figures 5.1 and 5.2 show not the actual count of total and censored posts I observed in the data but rather *estimated* posts and censored posts after applying a correction formula to extrapolate the total numbers of posts/censored posts generated prior to (unobserved) Sina censorship. See Appendix B for details about this formula.

Figure 5.2: Total and Censored Post Volume: Bo Scandal (Sep 17 - Dec 30)

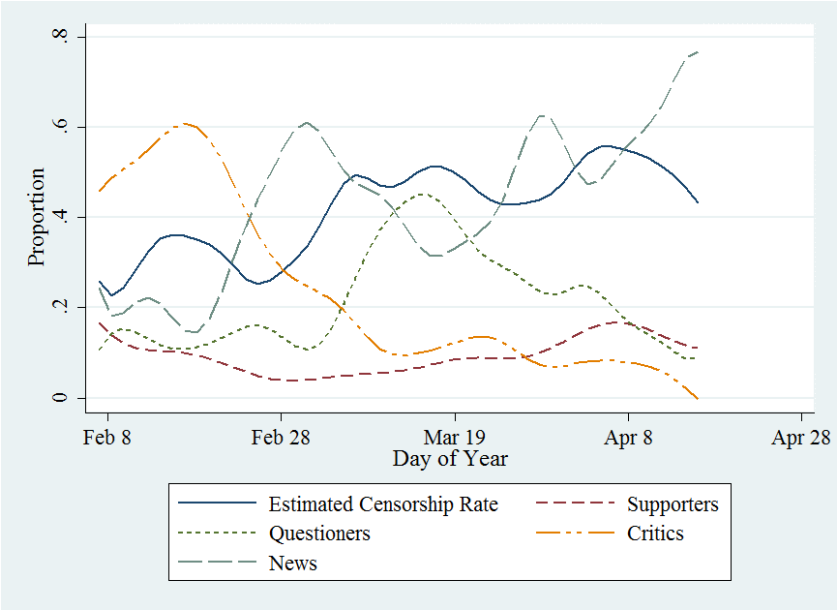


Three observations are immediately evident from the graph: censorship is higher, post peaks are lower, and overall volatility is higher with more peaks. While the censorship rate still varies, it is clearly above 50% of total posts for most of the phase, supporting my prediction that censorship would be highest during this time period. Along with this, post peaks are lower – this may be due both to an absence of major scandalous news (itself possibly due to controls on media reporting), and to higher censorship. Finally, volatility is higher. While I have no prior explanation for why this should be the case, it may be due to repeated and persistent state interventions into Bo-related discussion as well as the coincidence of the 18th Party Congress in November.

Next, I turn to examining fluctuations in the sentiment categories throughout Phases I-III. Figure 5.3 shows each category alongside the censorship rate in Phases I-II.[10]

[10]To generate these time series I used *ReadMe*, which estimated the daily category proportions

Figure 5.3: Sentiment Category Proportions: Bo Scandal (Feb 8 - Apr 17)

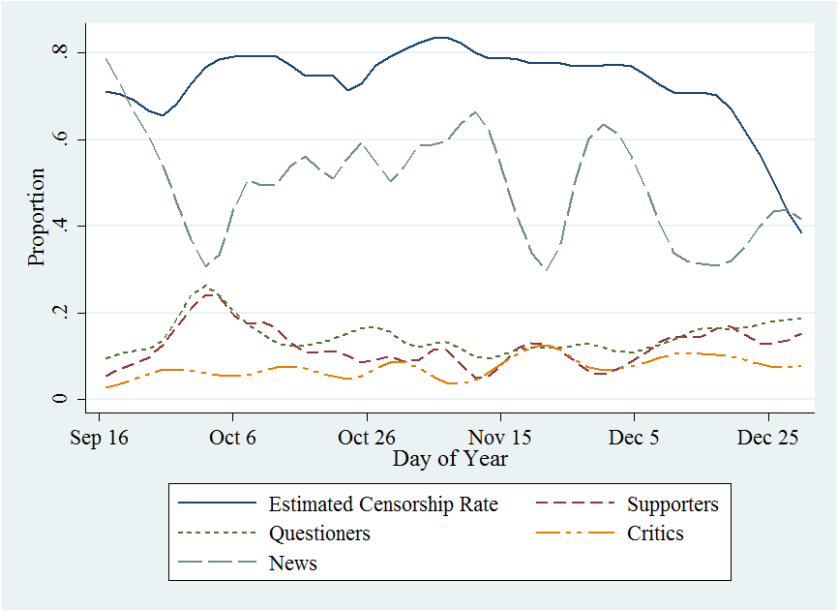


The graph's most notable aspect is the overall upward or downward trends of each series.[11] Censorship rises throughout, peaking near 60%. Its increase is mirrored in reverse by *Critics*, who peak early on during Phase I and then decline sharply. *Supporters* remains fairly low throughout but peaks in April during Phase II – the time when Bo came under the most sustained attack. Finally, *Questioners* peaks around March 14-16, a pivotal moment in which top leaders took action against Bo for the first time in removing him as Chongqing Secretary. Next, Figure 5.4 shows the trends in Phase III:

for *all* dates within each of Phases I-III (a total of 175 days), not just the narrower date ranges. I used the 1000 hand-coded posts as "training data" and then applied the *ReadMe* algorithm to the entire corpus of 68,885 posts. See Appendix A for details.

[11]Figures 5.3 and 5.4 display polynomial-smoothed time series to increase interpretability. I use the unsmoothed series in statistical analysis.

Figure 5.4: Sentiment Category Proportions: Bo Scandal (Sep 17 - Dec 30)



Censorship is obviously consistently higher than in Phases I-II, at least until the very end of the year. Meanwhile, *News*, while fluctuating, in general takes up a much larger proportion of content than previously. *Questioners* and *Supporters* are held to low levels and *Critics* is especially suppressed at around 10% of total content or less. Overall, Figure 5.4 is consistent with much higher censorship and overall tighter control (as evidenced by the substitution of blogger opinions for news content, much of it from state outlets) compared with the previous two phases.

The above two graphs give an initial idea of how the censorship rate relates to various sentiment categories. A more robust test, however, consists of examining their *dynamic* relationships: do increases or decreases in various categories affect the censorship rate (with some lag) in theoretically predicted ways? The next section considers this question.

5.4.1 Modeling the Sentiment Categories' Relation to Censorship

In this section I consider the statistical relationship between the category measures and the censorship rate, comparing Phases I-III. While I do not make specific predictions for each coefficient, in general I expect the directions and magnitudes of significant effects to resonate with Tables 5.1-5.3: in Phase I, lagged *Supporters* and *Questioners* measures should be negatively correlated with censorship and *Critics* positively correlated; in Phase II, *Questioners* should be negatively and *Critics* positively correlated; and in Phase III all three measures should positively correlate with censorship.

Similar to Chapter Four, I cannot use a standard linear model like OLS because of the likely violation of model assumptions. As before, I address these issues by employing Generalized Linear Model (GLM) regression and assuming that the censorship rate has a binomial distribution and that the model takes a logistic form. I then deal with autocorrelation by employing Newey-West standard errors. As in Chapter Four I used the AIC to select the model's maximum lag order, and in this chapter chose an order of two.

As previously, I incorporated this information into the model in two ways. First, I included the *observed measures* of lags 1-2 for all independent variables except censorship, and lags 1-4 (5 lags for Phase III) of censorship.[12] Second, I set the

[12]I dropped *News* for Tables 5.4 and 5.5 and kept it only in Table 5.6 for reasons of statistical power, since the former two tables had only 26 and 40 observations, respectively. Adding *News* up to two lags back into these regressions caused nonsensical results with extreme coefficients. A Variance Inflation Factor (VIF) test run on each set of Table 5.4 and Table 5.5 covariates (following estimation via OLS rather than GLM) revealed VIFs in excess of 100 for *News*, suggesting that it is highly collinear with the other sentiment categories and should be dropped, given the low N.

error term maximum lag to $N-2$; this should capture any residual interdependence among the series not accounted for by the included variables. Taking first Phase I, Table 5.5 shows average marginal effects:

Table 5.5: Sentiment Categories' Relation to the Censorship Rate (Phase I)

DV: Cens. Rate	Model I	Model II	Model III
L.Supporters	-0.527***	-0.482***	-0.333***
L2.Supporters		-0.513	-0.562*
L.Questioners	-0.066	-0.191**	-0.525**
L2.Questioners		-0.324**	-0.532***
L.Critics	0.058	0.087	0.176*
L2.Critics		0.093**	0.051
L.Cens. Rate			0.192
L2.Cens. Rate			-0.237**
L3.Cens. Rate			-0.385*
L4.Cens. Rate			0.247*

* $p < 0.1$ ** $p < .05$ *** $p < .01$ $N = 26$

As predicted, both *Supporters* and *Questioners* are consistently negative and significant across the fuller specifications (Models II and III). Moreover, as Table 5.1 theorized, *Questioners* has a greater negative magnitude than *Supporters*, consistent with the idea that commentary in the scandal's early weeks that merely attempted to find out what was going on and what connection Wang Lijun's flight might have to Bo was tolerated more than support for the Chongqing leader. However, I was surprised to note the relatively large magnitude at both lags one and two ($-.333$ and $-.562$) for *Supporters* as I had expected officials to tolerate some Bo commentary but not necessarily to open the floodgates. Lastly, lag one of *Critics* is positive and significant, but with a relatively small magnitude as predicted. Next, Table 5.6 gives Phase II results:

Results largely conform to prior predictions, but with one major exception. As predicted in Table 5.2, *Critics* is still positive and significant, and *Supporters* is

Table 5.6: Sentiment Categories' Relation to the Censorship Rate (Phase II)

DV: Cens. Rate	Model I	Model II	Model III
L.Supporters	-0.025	-0.013	0.017
L2.Supporters		0.046	-0.046
L.Questioners	0.095***	0.118***	0.174***
L2.Questioners		-0.024	-0.094**
L.Critics	0.085***	0.098***	0.100***
L2.Critics		-0.045	-0.087***
L.Cens. Rate			0.182***
L2.Cens. Rate			-0.091
L3.Cens. Rate			0.244***
L4.Cens. Rate			-0.059

$^{*}\ p < 0.1\ ^{**}\ p < .05\ ^{***}\ p < .01 \qquad N = 40$

near zero (and insignificant). The lag two results do diverge in sign from expectations, which does raise some concern since if an effect is really strong and persistent it should be consistent across both lags, but lag two is nonetheless less theoretically important than lag one because the former is further removed in time from the censorship rate's present behavior, weakening any inference (or lack thereof) about a causal linkage between the two. However, Table 5.6 does have one glaring anomaly: *Questioners* is 'wrong-signed' (positive rather than negative) and significant.

This finding does not necessarily undermine the overall value of *responsiveness benefit* and *credibility payoff* for the Bo case, but it may challenge the heavy weight (two times as strong as *visible censorship cost*) I gave these factors for Phase II in Table 5.2. In other words, top leaders may have decided that March and April, as opposed to February, were not the right time to relax censorship for signaling purposes with respect to *Questioners*. At the time, this category included numerous comments that while in my coding scheme were not explicitly opposed to the system or top leaders or even that skeptical, might have been construed by

leaders as dangerous simply because they asked too many probing questions about Bo's ties to other officials, or standing within the Party. That said, the most disciplined interpretation of this result is simply that it runs counter to theoretical expectations, without suggesting an alternative explanation.

I now turn to examining Phase III in Table 5.7:

Table 5.7: Sentiment Categories' Relation to the Censorship Rate (Phase III)

DV: Cens. Rate	Model I	Model II	Model III
L.Supporters	-0.046	-0.087*	-0.001
L2.Supporters		-0.028	0.024
L.Questioners	-0.118	0.018	0.061***
L2.Questioners		0.164***	0.227***
L.Critics	0.032	0.068	0.254***
L2.Critics		-0.081	0.087
L.Cens. Rate		0.442***	0.443***
L2.Cens. Rate		0.152***	0.166***
L3.Cens. Rate		0.142***	0.110***
L4.Cens. Rate		0.101	0.103
L5.Cens. Rate		0.160*	0.161**
L.News			0.074***
L2.News			0.068**

* $p < 0.1$ ** $p < .05$ *** $p < .01$ $N = 100$

Consistent with Table 5.3's predictions, both *Questioners* and *Critics* are positive and highly significant in most lags, indicating that each is associated with increased censorship. Additionally, while I had no strong theoretical prior as to *News'* coefficient, its positive sign and significance are consistent with prior work (Cairns and Plantan 2016) where a similar news measure also predicted higher censorship. Since much news is generated by state media sources, such an effect may not be direct but rather indirect: more news may prompt greater public commentary on an issue, which during times of high political sensitivity may be enough to trigger censors' response regardless of the commentary's specific content.

Just as with Table 5.6, there is one glaring anomaly in Table 5.7's results: *Supporters* is generally insignificant across specifications and near zero. Compared with the previous non-finding, however, this one is less of a concern because Figure 5.4 showed little if any activity at all in this category: there were simply very few *Supporters* posts in Phase III. Thus, the most straightforward interpretation is simply that the lack of variation in this category, combined with the still (relatively) small $N = 100$, led to either a genuine null finding or one due to weak statistical power. Either way, this non-result should not obscure the general conclusion that Table 5.7's coefficients are broadly consistent with prior predictions, and show a clear trend compared with Phases I-II toward multiple sentiment categories' relation to higher censorship.[13]

[13]Throughout analysis of Phases I-III I have not discussed the coefficients for lags one through four (or five for Phase III) of the censorship rate itself. This is because censorship is normally highly autoregressive, a result I have observed multiple times in prior work (Cairns and Carlson 2016; Cairns and Plantan 2016). Including censorship's own lags is vital for model specification, but itself does not have any theoretical importance beyond the fairly obvious statement that censorship tends to be 'sticky' over multiple dates as censors continuously purge social media of targeted content.

5.5 Conclusion: Progressively Higher Censorship From Phase I to III, But With Clear Cross-category Variation

In conclusion, I find that my theoretical predictions from Tables 5.1-5.3 are generally well-supported by the data. In general, censorship across all sentiment categories trends from looser to more stringent as one moves from Phase I to Phase III. The increasing share of news in Phase III as compared with I and II further reinforces the idea that as the year progressed and the Bo scandal unfolded, state censors became ever less willing to countenance organic speech on *Weibo*, instead allowing only the reposting of news and non-provocative commentary. This broad trend was not surprising given the high degree of political sensitivity leaders attribute to discussion of other elites' alleged misdeeds; in the framework's language, the *image harm* from allowing discussion is high. Yet even within overall tight control, the three categories diverged, with *Critics* being censored the most, *Questioners* less, and *Supporters* varying across phases. The fact that *Critics* predicted increased censorship across Tables 5.5-5.7 suggests, *pace* King, Pan and Roberts' work, that some sensitive topics *are* almost always censored even if they do not directly relate to collective action. While King, Pan and Roberts may be right in that 'ordinary' grumbling about the government is usually not censored, the sort of cynical and skeptical complaining that comprised much of the *Critics* category in this chapter is a clear exception: mere 'complaining', even when not an explicit call to arms against the state, may still not be allowed when it concerns embarrassing information about top elites themselves.

A second crucial result is that *Questioners* not only were not censored, but predicted *reduced* censorship in Phase I. While *Questioners* at first might appear to

be a less sensitive category than the other two, in reality mere factual information about some event is often what propels collective action (Kuran 1989; Lohmann 2002), as citizens generate common knowledge – they know that each other knows about some revelation, for example Bo's involvement in Neil Haywood's death. To be clear, shared information *per se* is not always sensitive, but in the case of elite-level official malfeasance we would expect any related information to be censored because "the government censors *all* posts in topics areas during volume bursts that discuss events with collective action potential. That is, the censors do not judge whether individual posts have collective action potential" (King, Pan and Roberts 2013, p. 7). In other words, King, Pan and Roberts have a binary conception of 'sensitive' versus 'non-sensitive' topics – the former have collective action potential while the latter do not. This dissertation's four-variable framework, in contrast, recognizes the existence of competing incentives within any given issue-moment, of which collective action is but one factor (albeit an important one).

As an alternative explanation, *Questioners* correlating with reduced censorship early on could be attributed to mere inaction on censors' part – in this account, *Weibo* attention to Wang Lijun's flight and connections to Bo caused a spike in posts while censors simply maintained their typical procedures of censoring only the obviously most extreme comments and did not crack down on the new posts – thus leading to a reduced overall censorship rate. This account would further maintain that censors did not adjust to the post surge because they had received no order(s) to do so, due to either top officials in Beijing simply being caught off-guard, or to genuine divisions about how to respond. While this explanation might be valid for a day, or perhaps at most a few days after Wang's flight, it cannot explain the persistence of lowered censorship in response to *Questioners* across the approximately 30 days of Phase I, given what we know about the normally swift

reaction times of the censors themselves and of supervising agencies.

In contrast, the four-variable framework offers two explanations. The first is that Chinese leaders allowed the public to raise questions about Wang and Bo in order to create common knowledge of Bo's guilt – authorities were betting that the public would interpret such obvious non-censorship as a credible sign that Beijing had abandoned him. Although Bo had his supporters, leaders likely felt that the majority of online citizens would turn against him once information about his misdeeds had become widely known. Supporting this interpretation, one individual I interviewed who worked at a high level within the Shanghai municipal propaganda system affirmed that the purpose of allowing such raucous speculation was to "ruin" or "smear" (Mandarin: *gao chou*) Bo.[14] Another prominent blogger also supported this theory, claiming that after authorities detained Wang subsequent to him leaving the U.S. Consulate, they allowed him to send a text message from his phone to journalists – in other words, to leak information that would eventually lead to tarnishing Bo's reputation.[15]

A second framework variable, *visible censorship cost*, accounts for why censorship *overall* was lower in Phase I than would be expected for such a sensitive issue. Indeed, this variable likely played a larger role during Phase I of this case than in any part of the other two studies in Chapters Four and Six. Once news of Wang's flight broke, it spread like wildfire, including mockery of the Chongqing government statement that Wang was receiving "vacation-style treatment" Had leaders chosen to censor the whole topic at this point, such a move would have only fed into already-rampant speculation about Wang's ties to Bo, and Bo's po-

[14]Interview by author, Shanghai, 12/13/14. While this individual was not directly involved in an agency with direct Internet oversight, his view provides a relative insider perspective on CCP media official thinking about what was unfolding during the scandal's early weeks.

[15]Source: PRI's The World. 4/11/2012. "China's Social Media Reacts Over Growing Political Scandal."

sition within the ranks of elite officials. Why, then, was *Questioners* positively rather than negatively signed in Phase II even though the scandal remained very visible? I had expected a more gradual transition where in state leaders' view, the news events of March and April would be a sort of transition period between the loose media environment of February, and the much tighter control they knew they would have to impose as the 18th Party Congress approached later in the year. The fact that openness regarding *Questioners* was relatively short-lived suggests that on highly sensitive issues, leaders are only prepared to tolerate small doses of common knowledge-generating speech even if they view benefits as outweighing risks in the short term. This then shows the limits of selective censorship during "hard" cases of online breaking events – even if elites occasionally permit lowered censorship, they view it as too explosive to persist for long.

5.5.1 Broader Implications and Future Research

The above findings have implications for multiple topics: the Bo case itself, President Xi's current anti-corruption efforts, and most broadly, for the possibility of observing selective censorship within regimes of overall tight authoritarian information control. Both for the Bo case and for Xi's broader anti-corruption campaign, the observed pattern of censorship and its relation to the sentiment categories supports the notion that some degree of transparency is *necessary* (rather than merely a good idea) for top leaders to obtain the support of China's largely educated, middle-income social media public in taking down particular officials. If leaders attempt to control public discussion following a major announcement too rigidly, the public may infer that the true state of affairs is worse than reported, since they think officials may be trying to hide bad news about (for instance)

their own involvement in the corruption, a claim echoed by recent formal models (Shadmehr and Bernhardt 2015). Yet the leadership's motives for occasionally loosening censorship may not be limited merely to preventing a negative inference by the public, but rather more positively persuading citizens that leaders are taking decisive action to combat high-level corruption. This seems to be a priority of the Xi administration, which has staked its legitimacy on disciplining wayward officials. Selective censorship for the purpose of gaining a *credibility payoff* is one tool central leaders can use in persuading a skeptical public to trust their ability to carry out anti-corruption efforts.

Finally, future research is needed as to the findings' generalizability both within China, and in similar authoritarian states that tightly restrict online media. It may be that the Bo case cannot be separated from its broader context in 2012 that affected the CCP leadership, namely the 18th Party Congress and leadership transition and the Party's overall vulnerability after decades of breakneck economic growth, environmental degradation, social ills and corruption. However, the subsequent investigation, expulsion from the CCP and trial of former Politburo Standing Committee member Zhou Yongkang, the highest-ranking official to be removed from power since 1989, instead supports the interpretation that Bo's case was not at all unique but a prototype for later takedowns of top officials. After the investigation against Zhou was made public on July 29, 2014, commenters on *Weibo* were observed to have considerable (temporary) freedom to criticize not only him, but the system as a whole.[16] While in this instance such openness was very short-lived and censorship quickly resumed, the fact that it occurred at all during an era of heightened media repression (under Xi) supports this paper's claim.

[16]Source: Wall Street Journal. 7/29/14. "Fall of Zhou Yongkang Lights Up China's Internet."

So far, Chapter Four has shown an issue where *credibility payoff* can be especially salient, while this chapter has also emphasized *visible censorship cost*. The next and last empirical chapter considers these factors as well, but uniquely among the dissertation's cases is an obvious instance of *collective action risk* in the context of the Diaoyu/Senkaku islands crisis.

CHAPTER 6

THE DIAOYU/SENKAKU ISLANDS DISPUTE AND 2012 DEMONSTRATIONS

6.1 Introduction

During summer and fall 2012 the ongoing Sino-Japanese conflict over the status of the Diaoyu/Senkaku islands escalated sharply. The crisis began in mid-August when a small group of protesters set sail from Hong Kong toward the disputed islands. Upon reaching their destination, they briefly landed on one of the islands but were then detained (although later released) by the Japanese, but not before taking pictures of the landing that went viral on Chinese social media. The detention enraged many Chinese, who viewed it as involving the unlawful arrest of fellow nationals on their own soil. Subsequently, on September 11 Tokyo made matters worse by purchasing part of the islands from private Japanese owners, in a move intended to de-escalate the standoff – since the Japanese owners themselves were viewed as hard-line nationalists – but was instead viewed by many in China as "nationalizing" the territory. This gesture catalyzed a series of mass Chinese protests against Japan, accompanied not only by angry anti-Japanese slogans, but also attacks on property and individuals perceived as having ties to Japan.

As with Chapters Four and Five, this chapter analyzes a major real-world incident that manifested prominently on *Weibo* as a means to evaluate Chapter Two's theoretical framework. To preview my method and findings, my co-author, an undergraduate research assistant, and I coded five sentiment categories that emerged as salient during the Diaoyu dispute's time frame (approximately

mid-August to late-September): *Moderate* commentary, *Patriotic* commentary, *Anti-Japanese* speech, *Calls to Action* (boycotts, sanctions or even military action against Japan), and lastly *Anti-Beijing* comments (that were directed at leaders in Beijing rather than the Japanese).[1] I find, in line with the logic of *responsiveness benefit*, that a sharp increase in anti-*Beijing* criticism during the initial protest wave in August was correlated with a *decrease* in censorship. Conversely, in September even moderate comments that actually condemned out-of-control protest violence in the streets, including the smashing of Japan-related businesses, were correlated with increased censorship – not because of their tone, but likely because they mentioned real-world collective action underway. Thus, while the case is a key example of *credibility payoff* in action, it ultimately shows the overriding weight of *collective action risk* in the state's calculus, especially during the sort of nationwide demonstrations that occurred.

Similar to previous chapters, the following sections analyze the Diaoyu dispute in Chapter Two's theoretical context by dividing protests into two phases. Phase I began on August 13 (just before the activist landing) and continued into early September, with rounds of street protests (and concurrent spikes in *Weibo* discussion) occurring over the weekends of August 18-19 and 25-26, and attention to the dispute slowly tapering off after that. Phase II kicked off on September 11 with Japan officially purchasing three islands, and peaked over the weekend of September 15-16 with massive street demonstrations and again on the 18th. Unlike in Chapters Four and Five, these two time periods were too brief (the N of days was too low) to undertake more complex statistical analysis. Instead, I rely more on descriptive statistics and graphs to show the relationship between various

[1]This chapter is based upon (and shares empirical results with) an article co-authored with Allen Carlson titled "Real-World Islands in a Social Media Sea"; see Cairns and Carlson (2016). However, the content of this chapter is strictly my own, and pertains to the dissertation's theoretical goals rather than the article's.

sentiment categories – especially *Anti-Beijing* speech, and *Moderate* comments – and the censorship rate during each Phase, and make only limited use of regression models. First, however, I briefly review the broader debate surrounding popular nationalism in China, its online form, and the state's role.

6.2 Relevant Literature

In studying the 2012 Diaoyu/Senkaku crisis, one key question is the extent to which nationalist sentiments on *Weibo* have exerted independent pressure on the Chinese state, versus being under their control. This debate has played out among scholars of Chinese politics and nationalism since the 1980s, with early contributions by Whiting (1983) and Oksenberg (1986), and more recent work by Zhao (2004) and Gries (2004). Collectively, these studies have shown how Chinese leaders have constructed nationalist narratives around a mythic dynastic past in which China was strong and glorious, followed by suffering at foreign hands during the "Century of Humiliation" and subsequent national resurrection under the CCP. However, scholars have divided on two fronts: first, the degree to which nationalist narratives have been primarily instrumental – inculcated by the state as means to secure popular loyalty to the CCP and to support elite goals – versus belief-driven, i.e. deeply rooted in popular consciousness and thus influencing or possibly even constraining leaders' menu of policy options in conducting foreign relations.

More recent work has tended to echo aspects of both these perspectives, portraying Chinese nationalism as a "double-edged sword" that leaders can both manipulate, and are sometimes beholden to (Hughes 2006; He 2007; Reilly 2011; Rozman 2013; Weiss 2013). For this dissertation, this more recent body of schol-

arship's major shortcoming is that it was mostly written prior to the rise of social media and particularly *Weibo*, or at any rate has not placed social media front and center as a crucial medium through which online nationalism is expressed. The 2012 Diaoyu/Senkaku confrontation was China's first international crisis in the *Weibo* era and one of *Weibo's* top ten trends of that year.[2] Such a situation alone calls for a detailed analysis of the Diaoyu case. The scope, energy and volatility of the 2012 protests dwarfed any other collective protest event in China that year, and was among the largest demonstrations ever to have taken place in mainland China. This markedly differentiates the case from Chapters Four and Five, neither of which involved any real-world demonstrations. Yet among the three chapters/cases, the Diaoyu episode perhaps also represents the most potent instance where leaders in Beijing might benefit from temporarily relaxing censorship in order to shore up their hard-line credentials with nationalist protesters, and conversely, an instance where where they might bear the greatest costs for prematurely suppressing dissent – even if aimed at Chinese leaders themselves. The Diaoyu crisis thus represents a "mixed" case for selective censorship and careful analysis of leaders' shifting costs and benefits at different time points and across specific sentiments is important for the theoretical framework.

Beyond the dissertation goals, however, the case also matters as evidence to help resolve the above debate about state-directed (instrumental) versus bottom-up (belief-driven) popular nationalism. While important theory-building has taken place about the balancing act Chinese leaders face in managing popular demonstrations, "weigh[ing] the risk to the status quo against the cost of using force or coercion to prevent citizens from gathering in the street" (Weiss 2013, p. 7), until

[2]Except for a more minor maritime confrontation in 2010 involving a Chinese fishing trawler; however, *Weibo* user numbers and the platform's overall "buzz" in 2010 were nowhere near the levels they reached in 2012.

recently authors have lacked robust empirical strategies to operationalize their concepts in online space and to develop measures of fluctuations in grassroots pressure in real-time.[3] The following analysis of *Weibo* data takes a step toward addressing this deficit by measuring which sentiments prevailed during the dispute, when they peaked, how they correlated with real-world events in the Diaoyu islands and mainland Chinese streets, and how the state responded with censorship.

6.3 Why the Diaoyu Dispute and How Was It Censored?

As mentioned above, the 2012 Diaoyu incident is a "mixed" case for selective censorship, with a greater degree of both "pull" (*responsiveness benefit* and *visible censorship cost*) and "push" (*image harm* and *collective action risk*) factors for lowering *Weibo* censorship present during the dispute than in either of the previous two chapters. Thus, it does not make sense to make a single, overall prediction of high or low censorship for the case; rather, I expect censorship to vary markedly across the two phases, and across different sentiments within each phase.

A more detailed summary of the course of events provides the context for this distinction. Beginning in August, the activists were detained on the 15th and released two days later. Initial reports of street protests in China broke on August 15.[4] On August 17, Japan's Kyodo news service reported that messages had appeared on *Weibo* calling for anti-Japan demonstrations, and noted that the posts were not immediately deleted.[5] On August 19, a group of Japanese activists

[3]Notable exceptions include Shen 2007; Callahan 2009; Johnston and Stockmann 2007; Carlson 2009; Gries et al. 2011; Cheng 2011; Hoffman and Lerner 2013.

[4]Source: Kyodo News Service (reprinted by BBC Monitoring Asia Pacific. August 15, 2012. "Chinese group protests in front of Japan embassy to demand disputed isles".

[5]Source: Kyodo News Service (reprinted by BBC Monitoring Asia Pacific. August 17, 2012.

landed on the islands. Perhaps anticipating, or at any rate quickly reacting to this development, protests unfolded that same day in several major Chinese cities and were covered by Xinhua.[6] By the 21st, however, mainstream media were attempting to calm protesters and avert further violent escalation.[7]

In September, the government took a different approach to managing the protests in response to Japan's island purchase, tolerating or encouraging rather than de-escalating them. Leaders continued to tolerate massive street demonstrations across mainland China through September 18. Moreover, Chinese Foreign Ministry spokesman Hong Lei explicitly mentioned the anti-Japan demonstrations as late as September 19 as justified to defend Chinese sovereignty, while Xi Jinping called the islands' purchase a "farce".[8] The *People's Daily* and other major Chinese media outlets also published commentaries, which ranged from "sympathy" to outright support for the demonstrations.[9]

Based on a reading of officials' response to the protests as well as more theoretical consideration of the August and September waves, I expect Phase I to have lower censorship than Phase II, for four reasons. First, generally speaking, the "pros" of lower censorship tend to be strongest towards the beginning or middle of online breaking incidents, with the "cons" increasingly winning out as crises drag on. Second, while Phase I did witness substantial street protests, Phase II in September had truly massive collective action involving hundreds of thousands of

"Internet messages in China call for anti-Japan protests on Sunday". That same day, Kyodo reported that protesters in Shenzhen smashed Japanese-branded cars and broke into Japanese restaurants.

[6]Source: Xinhua (reprinted by BBC Monitoring Asia Pacific. August 19, 2012. "Protests in China against Japanese activists' visit to disputed islands".

[7]Source: South China Morning Post. August 21, 2012. "Bid to calm public after ugly Diaoyu's protest; Media praise patriotism over disputed islands, but some call Shenzhen behavior 'shameful'".

[8]Source: China Daily European Edition. September 20, 2012. "Xi slams Diaoyu 'purchase'".

[9]Source: The New York Times. September 17, 2012. "Beijing Mixes Messages Over Anti-Japan Protests".

demonstrators.

Third, the nature of the specific real-world provocation differed. In Phase I, the Diaoyu activists' detention was the proximate cause of online and street protests and criticisms. While news reports had mentioned Japan's plan to "nationalize" some of the islands by then, it was not clear if or when the Japanese legislature would actually formalize the purchase. Phase II, in contrast, was triggered by the islands' purchase actually being approved, an even more provocative move (in the eyes of many Chinese) as it invoked deeper questions about territorial jurisdiction, and by extension sovereignty. Fourth and finally, Chinese leaders themselves took a harder-line stance in Phase II compared with Phase I, perhaps having held back their most intense criticisms of Japan in August in the hope that Japan would not move forward with the purchase. While leaders' hard line might seem an invitation for *Weibo* commentators to freely express their own extreme sentiments, the four-variable framework would suggest the opposite: *responsiveness benefit* exists as a means to persuade citizens that the government will implement some *future* action. Once officials actually shift approach (or in this case, make statements defending China's territorial claims), the government's need to lower censorship is reduced or eliminated.

6.3.1 Explaining Censorship Variation Across the *Moderates*, *Patriotic*, *Calls to Action* and *Anti-Beijing* Sentiment Categories

Just as in the previous two chapters, I operationalize Chapter Two's four-

variable 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*, and here add *collective action risk* to the analysis for the first time. As before, I 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. I also assume that *collective action risk* varies across both categories, and time.

Then, in Tables 6.1-6.2 below I score *credibility payoff* and *visible censorship cost* on a scale from -2 to +2, with -2 as "Very Negative", 0 as "Neutral" and +2 as "Very Positive". As before, I weight *credibility payoff* as twice as important (2x) as *visible censorship cost*, and then sum the two scores. However, unlike Chapters Four and Five, the addition of *collective action risk* changes the end equation. I conceptualize this new factor as *deterministically* predicting high censorship above a certain threshold. I code it as either "Negative" (-1), "Neutral" (0) or "Positive" ($+1$). If it is Negative, then the prediction equation simplifies to that found in Chapters Four and Five, involving only *credibility payoff* and *visible censorship cost*. If *collective action risk* is "Positive", however, then the state *always* seeks to achieve maximum censorship of those sentiment categories that bear the "Positive" label. In the case where it is "Neutral" I strike a balance by setting its score as +6, and adding this constant to the equation. That is, the weight of "Neutral" *collective action risk* is equal to the weight of both "Very Positive" *credibility payoff* and *visible censorship cost* combined. With "Neutral" collective risk, the lowest that censorship can be for a given category/moment is "partial" or medium censorship. The following equation formally defines these relationships:

$$Cens = \begin{cases} +6 & \text{if } ColActnRisk > 0 \\ -2 * Credibilitypayoff + -1 * VisCensCost + 6 & \text{if } ColActnRisk = 0 \\ -2 * Credibilitypayoff + -1 * VisCensCost & \text{otherwise} \end{cases}$$

With this equation in hand, I can now make the predictions in Tables 6.1 and 6.2:[10]

Table 6.1: Predicted Censorship by Sentiment Category (Phase I: Aug. 13 - Sept. 10)

Category	Cred. Payoff ($-2x$)	Vis. Cens. Cost ($-x$)	Col. Actn. Risk (threshold)	Pred. Censorship
Moderates	N/A	N/A	N/A	N/A
Patriotic	Neutral (0)	V. Positive (+2)	Negative (-1)	-2
Action Calls	Positive (+1)	V. Positive (+2)	Neutral (0)	+2
Anti-Beijing	V. Positive (+2)	V. Positive (+2)	Negative (-1)	-6

Table 6.2: Predicted Censorship by Sentiment Category (Phase II: Sept. 11-30)

Category	Cred. Payoff ($-2x$)	Vis. Cens. Cost ($-x$)	Col. Actn. Risk (threshold)	Pred. Censorship
Moderates	Neutral (0)	Positive (+1)	Positive (+1)	+6
Patriotic	Positive (+1)	Positive (+1)	Negative (-1)	-3
Action Calls	Negative (-1)	Positive (+1)	Positive (+1)	+6
Anti-Beijing	Negative (-1)	Positive (+1)	Neutral (0)	+6

[10] *Moderates* is absent from Table 6.1 because I found this category to simply be too sparse in Phase I to include it in the predictions. Based both on a reading of the other posts and knowledge of real-world protest events in August, I do not think that the reason it is sparse is due to censorship, but rather the mere lack of a vocal moderate faction online at that time.

I scale censorship from -6 to +6, with -6 representing very low censorship, +6 near-total or total censorship, and 0 partial censorship.[11]

Beginning with Phase I, I coded *visible censorship cost* as Very Positive because the activist landing and detention was widely reported (and not censored) within China, providing a focal point around which protesters could mobilize. Chinese news media also did much to play up the incident, which would have made sudden online censorship very obvious. I coded *collective action risk* as either Negative or Neutral because although street protests did immediately take place, it was not at all clear in August that *Weibo* discussion was adding fuel to street actions, or triggering a broader collective action cascade. *Calls to Action* were a partial exception in that some demanded boycott actions against Japanese products. For *credibility payoff* I coded *Patriotic* calls as Neutral, *Calls to Action* as Positive and *Anti-Beijing* sentiments as Very Positive, and elaborate more on each category's general nature in the next section. Here, I note that the latter two categories scored as +1 or +2 precisely because of their provocative demand-making vis-à-vis the state, with *Anti-Beijing* sentiments more so in that they explicitly criticized the government. And compared with Phase II, the downsides (potential for *image harm*) of each were not as salient, because such calls had only just begun to appear and had not morphed into a much broader anti-government movement.

For Phase II, I coded *visible censorship cost* as Positive: obviously, the huge amount of *Weibo*, street, and print media activity meant that imposing firm cen-

[11] As in Chapters Four and Five, I do not intend to predict actual censorship levels but rather develop a categorical scheme to capture the variables' relationship with censorship's *relative* magnitude. Mathematically, the highest theoretical value should be +12, which would be the case if $ColActnRisk = 0$ and both *credibility payoff* and *visible censorship cost* were -2. However, for practical purposes, values greater than +6 are irrelevant. The point of the $if\ ColActnRisk = 0$ condition in the above equation is to show the balance of censorship factors possible if collective risk is offset by a positive value for the other two variables. If they are negative, then the situation is essentially similar to when $ColActnRisk > 0$.

sorship (or shutting down the street protests, for that matter) would be highly visible. However, it would not be as *conspicuous* as in August; that is, most Chinese would already have "priced in" the expectation that their leaders were surely going to shut down protests and discussion sometime soon. Thus, even though highly visible, the imposition of censorship, particularly after the raucous weekend of September 15-16, would likely not be seen as unusual. For *collective action risk*, I coded *Moderate* voices as Positive mainly because as I discuss below, they contained references to actual protest events. *Patriotic* slogans were the one category with Negative *collective action risk* in September: such sentiments were so widespread on *Weibo* and in society (and echoed by state-supervised media everywhere) as to be innocuous. Finally, *Calls to Action* and *Anti-Beijing* sentiments were Positive and Neutral, respectively, because each carried a greater risk of adding fuel to the protest fire compared with August, since the protests themselves (and protesters' energy level) had grown so substantially. Finally, I generally coded *credibility payoff* for each category as more negative than before due to the much greater overall political sensitivity of September's events; *Patriotic* speech was again the one exception to this.

As in previous chapters, I next study dynamic trends in the censorship rate and the sentiment categories as observable implications of Tables 6.1-6.2. First, though, I address how my co-author and I isolated posts relevant to the Diaoyu protests from the WeiboScope data, and how we defined and coded the sentiment categories.

6.3.2 Identifying Discussion of the Dispute and Coding the Sentiment Categories

To prepare the data for analysis, my co-author and I extracted from the WeiboScope data those posts from August 13 – September 30, as we found these to contain the vast majority of Diaoyu-related commentary. Next we filtered the posts according to whether they contained a keyword, "Diaoyu Islands" (*diaoyudao*), that was highly predictive of discussion of the dispute. This left about 145,000 posts over 49 days. One immediate concern with filtering on this (or any) keyword is whether estimates of sentiment categories derived from such a sample are generalizable to the broader population of all topic-relevant posts. This concern is heightened compared with Chapters Four and Five, which used a broad basket of keywords to sample their respective topics. Accordingly, I show results below for both posts containing "Diaoyu Islands", and a sample that did not but which I found, through careful reading, to also be topically relevant. To preview my findings, adding this additional sample strengthens the overall results.

I now describe the coding procedure in greater detail. My co-author, our undergraduate assistant, and I worked to assign a random sample of 479 posts into one of eight sentiment categories, of which *Moderates*, *Patriotic*, *Calls to Action*, and *Anti-Beijing* ultimately emerged as the most theoretically important. After working independently, the three coders met to reconcile scores, and we report inter-coder reliability statistics in Appendix A. We defined *Moderates* as posts that cautioned against the potential negative consequences of direct action against Japan, or objected to the violent turn that protests took over the course of the 49-day sample. The latter of these views is exemplified by the statement, "Pro-

tecting the Diaoyu Islands does not involve harming our fellow citizens!" (*baowei diaoyudao bushi shanghai ziji de tongbao*). As such posts mostly opposed the virulent and even violent excesses of more extreme fellow Chinese, we initially felt that Beijing would view these voices as posing little risk. However, a closer inspection of the posts, particularly in September, revealed that many contained references to actual protest events. Per King, Pan and Roberts' (2013; 2014) understanding of "collective action potential", such posts should be censored "regardless of content" (p. 1). This is what led to this category's eventual designation as posing "High" collective action risk, particularly in September when such posts appeared alongside real-world demonstrations.

The second category, *Patriotic*, entailed posters who parroted Beijing's own official statements about the Diaoyu Islands, in particular their status as part of China, and the country's determination not to cede such territory. While such a category could encompass a wide variety of phrases, we anticipated its most common refrain to simply be the patriotic assertion, "The Diaoyu Islands are China's" (*diaoyudao shi zhongguode*). While such slogans often manifested during the street demonstrations and thus the category might appear to involve collective action risk, these statements were not coterminous with actual demonstrations but rather widespread throughout *Weibo*, and were very orthodox vis-à-vis the state. Thus, Chinese leaders could expect to receive little *responsiveness benefit* from allowing this category to proliferate, but did stand to suffer substantial *visible censorship cost* if they suppressed such loyally and obviously patriotic sentiments.

A third category encompassed posts containing strongly anti-Japanese sentiments, and those that actively blamed the Chinese people (not state) for their weakness in not standing up to Japan. The "anti-Japanese" subset of this cate-

gory generally spouted derogatory anti-Japanese rhetoric, such as "little Japs" (*xiaoriben*), "Japanese pirates" (*wokou*), or "Japanese devils" (*riguizi*). In contrast, the second subset entailed lamenting China's inferiority or fecklessness vis-à-vis Japan without in any way implicating Beijing. While these two subsets are substantively very different, we estimated that both posed only a moderate degree of risk to the state. Ultimately, I could not reconcile this category with the prediction exercise in Tables 6.1 and 6.2 since it occupied a murky "middle ground" between other, better delineated statements with respect to the state's risk/benefit; for this reason, it was dropped from the analysis.

A fourth category consisted of posts containing specific *Calls to Action*, be they endorsements of "boycotting Japanese goods" (*dizhi rihuo*) , support for "protests" (*kangyi*), or even demands to "deploy military forces" (*pai bing*). Not surprisingly, we found this category to be quite heterogeneous. Such sentiments went well beyond the official actions that Beijing had taken, and were more pointed than the other categories as they contained specific suggestions and/or threats regarding how best to settle China's score with Japan. As such, these posts increased collective action risk for the government, especially since such calls often coincided with similar slogans and street actions, but could also be viewed as an opportunity for Beijing to appease the ultra-nationalist constituency within Chinese society and to gain *responsiveness benefit*.

The fifth category, *Anti-Beijing*, entailed posts that criticized the Chinese government in some manner, whether by referring to the abuses of power of China's notorious "City Urban Administrative and Law Enforcement Bureau" (*chengguan*) — cemented in the popular consciousness as low-level police "thugs" – highlighting the impotence of China's "dear military" (*guibing*) or of the Communist Party, or

more generically criticizing the Chinese state. While such posts might appear at first to pose the greatest risk to the government, in reality the four-variable framework would classify them more moderately. They do not pose a collective action risk *per se*, nor do they harm Beijing's image as long as the anti-government statements consist of anger at Beijing on *nationalist* issues – which my co-author and I found to be most frequent – rather than over other domestic issues like corruption, and as long as leaders are willing to signal through words and actions that they are willing to stand up to Japan. On the other hand, anti-Beijing criticism could potentially yield a large *responsiveness benefit*, as per the logic discussed in the previous section.

Last, to complete the category scheme, my co-author and I identified three additional categories of Diaoyu-relevant content: *News*, *Humor*, and *Other*. *News* consisted of posts that either consisted of, or contained a re-tweet of a print media story or opinion piece, and that did not fit under any of the other categories (which all required the expression or re-posting of an original microblogger sentiment). This category mattered as a control variable since it captured the circulation of real-world news on *Weibo*, but was otherwise not theoretically important. *Humor* involved jokes, satire, puns etc. that did not otherwise fall under one of the more substantive categories – some humor was clearly not a mere joke but intended to send a political message, and if so we classified it elsewhere. Finally, *Other* was a residual category to capture both irrelevant posts to the Diaoyu topic, and occasional relevant posts that our coding team could not agree to classify under any of the other categories.

As with Chapters Four and Five, my end goal was twofold: to directly estimate category proportions on key dates as evidence to evaluate the predictions in Tables

6.1-6.2, and to create time series measures of the categories across the entire 49-day dispute. Toward this first objective, my co-author, our undergraduate assistant and I drew additional samples of 150 posts each from four dates of peak *Weibo* activity that also corresponded to real-world events: August 16 and 18, and September 11 and 15. August 16 was the day after the activists landed on the islands and a date where they remained in Japanese detention; August 18 was the day after their release, but right before a Japanese nationalist group sent its own landing party; September 11 was the date of the islands' purchase; and September 15 was the beginning of the weekend that saw the largest protests of any weekend during the whole dispute – all dates were chosen to be representative of Phases I or II, respectively. We then individually coded the posts and as before, met to reconcile scores, the results of which appear in the next section.[12]

In addition to the directly estimated proportions, and *ReadMe* estimation (which was done using the original 479 post sample), we also relied on keyword measures to proxy for the sentiment categories, which I describe here.[13] Beginning with *Moderates*, we discovered that two terms, "rational patriotism" (*lixing aiguo*), and "smash" (*za*, – with reference to admonishments not to carry out such activities against Japanese goods owned by Chinese citizens), were the most prevalent terms. Not surprisingly, the simple declaration that "The Diaoyu Islands are China's" was the dominant refrain within the *Patriotic* category. Next, the *Call to Action* category was dominated by various statements in support of the "boycotts of Japanese goods" then underway within China during late summer and early fall 2012. Finally, the *Anti-Beijing* category, while containing a wide array of criticisms, was most consistently voiced via protesters' satirical use of the

[12]Our procedure was identical to that followed in the first coding round, and explained in Appendix A.

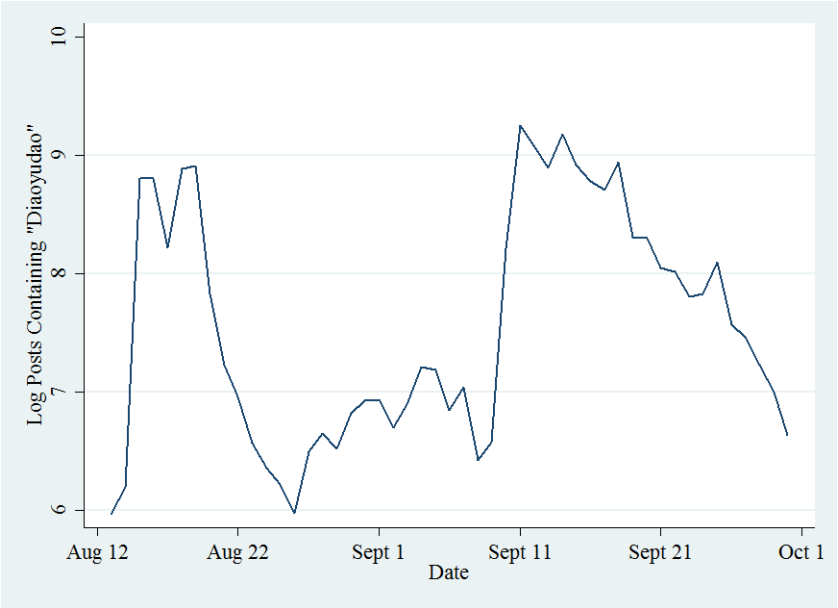
[13]Details about *ReadMe's* application to this chapter are in Appendix C.

phrase "heavenly dynasty" (*tianchao*), a reference to the ineptitude and decline of the Qing dynasty, to refer to the contemporary Chinese state. While on the surface the phrase is one of deep respect and admiration, within the context of the posts it was used to underscore the shortcomings of China's leaders and the domestic political status quo. My co-author and I derived this read of the term from the manner in which we so consistently saw it used in unison with sharp criticism of government policies, and also reviewing the secondary literature, which has recently highlighted the term as a particularly biting form of anti-government commentary (Link and Xiao 2013).

6.4 Results

Before examining the sentiment categories, it is useful to begin the analysis by looking at post volume throughout the dispute, displayed in Figure 6.1:

Figure 6.1: Volume of Posts Containing "Diaoyu Islands" (Aug 13 - Sep 30)



The two waves that comprise Phase I and Phase II are evident, with peaks on August 16 and 18, and September 11 and 15. In the absence of regression analysis, empirical results in this chapter depend on observing spikes in sentiment categories predicted to yield high or low censorship alongside graphically noticeable changes in censorship either occurring simultaneously with, or immediately following these changes. As a first check on whether this obtains, I focus on two key dates of especially volatile *Weibo* activity: August 18 and September 15. August 18 witnessed a surge in *Anti-Beijing* comments, while September 15, though not without anti-government sentiments, saw an uptick in *Moderate* posts. Table 6.3 shows estimated censorship rates for these two dates, both among posts containing *diaoyudao* and posts without this keyword:[14]

[14]As a robustness check that the data sample was not prone to "diaoyudao" keyword selection bias, I analyzed a small sample of posts without the keyword. The estimates are thus subject to some sampling variance. For August 18, I drew a sample of 49 posts, of which none were censored. As no "successes" occurred in this binomial variable, I applied Hanley and Lippman-Hand's "Rule of Three" (1983) to estimate the upper 95% confidence bound. For September

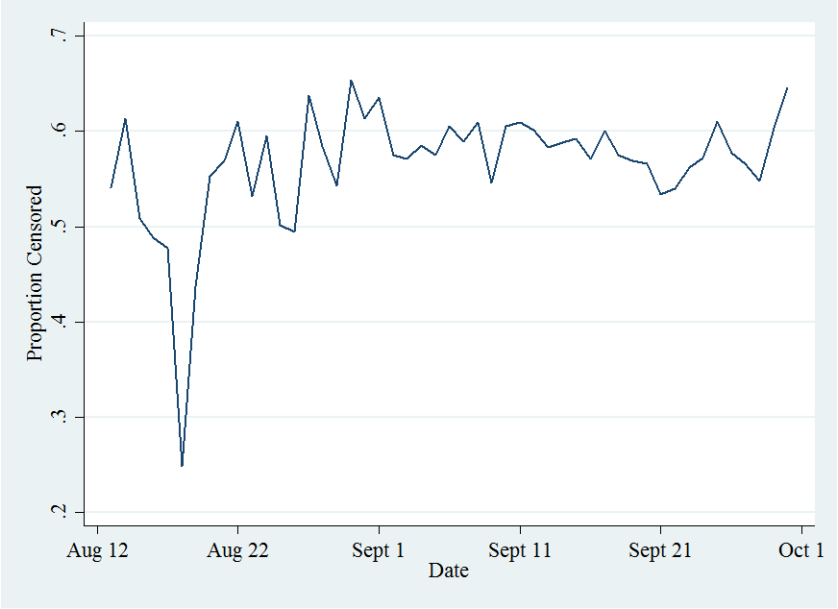
Table 6.3: Censorship Rates during Dispute Discussion Peaks With and Without the *Diaoyudao* Keyword

Date	95% Confidence Interval (mean): posts without *diaoyudao*	95% Confidence Interval (mean): posts with *diaoyudao*
August 18	0-38.5% (0%)	22.3-27.1% (24.7%)
September 15	51.0-82.5% (69.9%)	58.9-62.1% (60.5%)

While confidence intervals for sample estimates are wider due to the smaller sample size of non-*diaoyudao* keyword posts, it is evident that censorship is much lower on August 18 than September 15; in fact, the difference may be even greater than for keyword posts, as the lower bound on August 18 for the non-keyword sample is at zero and the upper bound for September 15 is as high as 82.5%. As a next step, I estimate the censorship rate (for posts containing *diaoyudao*) across the entire time period, shown in Figure 6.2 below:

18, I sampled 51 posts, of which 10 were censored, and calculated the confidence interval from a binomial distribution. In both cases, I adjusted the observed censorship rates in the WeiboScope data to account for the data's downward bias. The numbers here rely on Zhu et al.'s (2013) finding that 90% of post deletions occur within 24 hours of an initial event. I assume the data have the same speed of censorship as theirs. See Appendix B for more details.

Figure 6.2: Proportion of Posts Containing "Diaoyu Islands" That Were Censored (Aug 13 - Sep 30)



This graph shows that the censorship rate remained fairly constant throughout the dispute except on August 18, when it plummeted to about 25%. This contrast suggests that authorities' ordinary response for much of the dispute was to censor at a high rate – the fact that the rate dropped so sharply in mid-August, therefore, suggests that they made a deliberate decision to censor less at that time.

I next estimate sample proportions for all categories both averaged across the entire time series, and for specific days. The 49-day results were estimated directly from the hand coding of 479 observations, and the day-specific results from the 150 observations drawn for each of the four targeted days, displayed in Table 6.4 along with the censorship rate:[15]

[15]This rate uses the correction formula from Appendix B rather than the raw (unadjusted) figure.

Table 6.4: Sentiment Category Proportions and the Censorship Rate during Peaks in Diaoyu Discussion

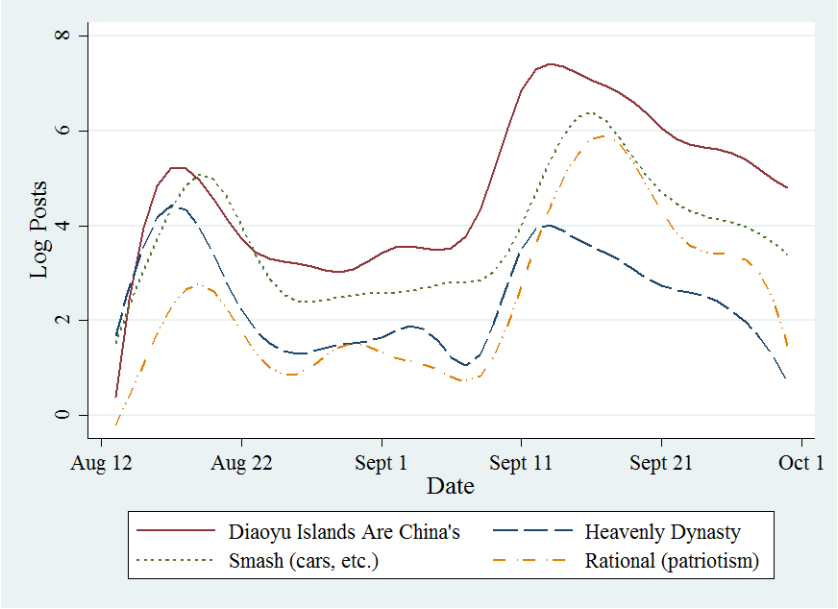
Category (% posts)	8/13-9/30 (avg.)	8/16	8/18	9/11	9/15
Moderate	4	6	2	3	19
Patriotic	13	13	6	21	9
Call to action	9	11	21	11	6
Anti-government	13	29	51	7	10
Censorship rate	57	49	25	61	59

The difference between the two August, and two September dates is immediately evident. The censorship rate is lower than average on August 16 and 18, which are the dates in which *Calls to Action* and *Anti-Beijing* sentiments predominate. In contrast, *Moderates* reach their highest point of the year on September 15, when the censorship rate is above average. As single dates, these observations are roughly consistent with the theoretical predictions in Tables 6.1 and 6.2, although by themselves they are certainly not conclusive.

While these direct estimates give a general idea of sentiment category fluctuations, graphing the ebb and flow of keywords representative of these categories is another useful way to relate them to censorship. Figure 6.3 shows these for select keywords:[16]

[16]To ensure topic relevance, we constructed these series only from posts containing *diaoyudao*. The graphs are smoothed with second-order polynomials.

Figure 6.3: Log Volume of Posts Containing Sentiment Category Keywords: Diaoyu Dispute



The graph's most distinctive aspect is the sequencing of the four time series. The "Diaoyu Islands Are China's" patriotic refrain, as Sina *Weibo's* tenth-most common 'hot topic' (*remen huati*) of 2012, proxies for micro-bloggers' general attention to the dispute, and particularly their response to real-world events – this phrase surges immediately following the activist landing on August 15, and actually anticipates by a few days Japan's purchase of the islands on September 11.[17] The other keyword series then unfold in the context of this macro trend. "Heavenly Dynasty" (*tianchao*) peaks on August 18 during the moment of lowest censorship but also shows an uptick in September, while two phrases representing a moderate backlash – decrying the "smash[ing]" of Japanese cars, and calling for "rational" patriotic expression – surge in mid-September amid relatively high censorship (close to 60%).

[17]Source: *Weibo* data center: http://tech.sina.com.cn/i/2012-12-19/13447902817.shtml.

6.4.1 Modeling the Sentiment Categories' Relation to Censorship

Although as mentioned earlier, the number of observations in each phase was too low to rely solely on regression results, I did run the regressions of the censorship rate on the *Moderate*, *Patriotic*, *Calls to Action* and *Anti-Beijing* predictors with only two control variables included: lag one of censorship, and of the predictor in question. Each sentiment category time series was *ReadMe*-generated in a similar manner to the previous chapters.[18] Like Chapters 4-5, the regressions are generalized linear models with Newey-West standard errors and a binomial distribution. Table 6.5 shows results from these pairwise model runs in Phases I and II:

Table 6.5: Sentiment Categories' Relation to the Censorship Rate: Pairwise Regressions (Phases I and II)

DV: Cens. Rate	Phase I	Phase II
L.Cens Rate	0.413***	0.438**
Moderates	-0.010	0.083**
L.Moderates	0.154	-0.271***
L.Cens Rate	0.460***	0.441***
Action Calls	0.212*	0.118***
L.Action Calls	-0.229	0.110*
L.Cens Rate	-0.013	0.499***
Anti-Beijing	-0.219***	-0.072***
L.Anti-Beijing	-0.363	0.067

$^{*}\ p < 0.1\ ^{**}\ p < .05\ ^{***}\ p < .01 \qquad N = 27\ (PhaseI);\ 21\ (PhaseII)$

[18] As in Chapters Four and Five, examination of the censorship rate's autocorrelation and partial autocorrelation graphs showed it to be autoregressive; in the Diaoyu case only lag one is significant. Lag one of the other predictors for this case was also autoregressive.

The sign and significance of coefficients is generally consistent with theoretical predictions with some exceptions. *Moderates* at lag zero is insignificant in Phase I but positive and significant in Phase II. *Anti-Beijing* lag zero is negative and significant in Phase I, predicting lower censorship which is consistent with Table 6.1. Surprisingly, it is also negative and significant in Phase II, although with a much smaller magnitude. Further comparison between the two phases is needed as to the nature of anti-Beijing sentiments, but it may be that they posed less of a risk of *image harm* to Beijing's reputation in Phase II than I had originally estimated, even if the *responsiveness benefit* of allowing them was no longer as potent. And *Calls to Action* is positive and significant in both phases (except for lag one in Phase I), suggesting that this category of speech was of little benefit to leaders, but sensitive as it contained slogans like "boycott Japanese goods!" that also mobilized demonstrators in the real-world.

Of course, these correlations need to be interpreted very cautiously given the low N and the lack of controls in each pairwise regression for other sentiment categories' behavior. By themselves, they are clearly insufficient to evaluate the theoretical predictions. Yet when viewed in the context of the overall variation in sentiment categories and the censorship rate across the two phases, they support the general argument (if not each specific prediction) of Tables 6.1-6.2: higher censorship in Table 6.2 compared with Table 6.1, and a close correlation between the behavior of *Anti-Beijing* sentiments and the censorship rate. The positive sign (and significance) for *Calls to Action* in Phase II compared with Phase I also resonates with the predictions.

6.5 Conclusion: Lowered Censorship in August During *Anti-Beijing* Comments; High Across-the-board Censorship in September

In conclusion, the previous section's results highlight the importance of *collective action risk* in shaping what topics get censored (and when) during high-profile online incidents like the Diaoyu protests. Censorship showed less variation overall than in Chapters Four and Five, remaining high throughout most of the 49-day dispute. The one exception to this was a major dip around August 18, which when combined with the surging presence of both *Anti-Beijing* commentary and *Calls to Action* on that date, poses a major challenge to overly simplistic understandings of *collective action risk* that treat it as deterministically causing higher censorship. King, Pan and Roberts (2013) specifically mention nationalist protests (p. 6) as especially likely to prompt censorship due to their strong association with real-world collective action. If we take the authors at their word, then the clear drop in censorship on August 18 is more than a mere caveat to their findings, but potentially more broadly calls into question their study's appropriate scope. This is not to neglect their major contribution in defining collective action potential and showing a wide range of clear instances of on-the-ground protest where censorship surges, but it does suggest that the phenomenon of social media censorship during breaking incidents in particular is better explained by a multi-causal framework like the one in this dissertation.

Regarding censorship and *Weibo* activity in September, the picture is less crisp, but at a minimum, the finding that even *Moderate* comments were censored further complicates the simplistic application of *collective action risk*. However, Septem-

ber's events are murky with respect to *credibility payoff*: the logic that the government needed to convince online citizens of its resolve was certainly less salient in September than in August, and I thought that this difference would equally apply to all sentiment categories. Instead, *Anti-Beijing* comments were correlated with lower censorship in September as well, which raises the possibility that this particular sentiment category may be surprisingly immune from the censors during incidents where central leader credibility is at stake. September's implications for *visible censorship cost* also require further investigation. While the street protests and associated online commentary were highly visible to the public in both August and September, the Diaoyu case points to a need to further refine understanding of this variable by looking at its *conspicuousness* at any given time point. That is, future studies should look not only at how visible an online sentiment is at any moment, but what netizen expectations are regarding how "acceptable" suddenly deleting it would be. During particularly unstable times, netizens may view the deletion of provocative sentiments (or those that obviously incite collective action) as not only unremarkable or expected, but a legitimate decision to preserve social stability, and will not draw a negative inference about a government cover-up from their disappearance.

Overall, the Diaoyu case shows the empirical difficulties inherent in using short-lived, acute crises to study selective censorship. The four-variable framework is best tested with multiple "observations" (such as daily measures of sentiment category proportions) over time in relation to censorship. When incidents or incident phases last only a week or so, applying statistical methods is very difficult. One solution could be to measure sentiments and censorship at a resolution of hours rather than days, but this would require finer-grained controls to account for daily cycles in the censors' behavior – a possibility I address in Appendix B.1. That said,

even if such brief cases are difficult, they are critically important for untangling interactions among the four variables when all four are active at once, particularly the interplay between *collective action risk* during acute crises and the other three factors. Although *collective action risk* often trumps the other variables when present, the case suggests that even during the early stages of what authorities had to have known would be substantial street protests in August, leaders might be willing to tolerate some protest risk if justified by other potential strategic gains.

6.5.1 Broader Implications and Future Research

Beyond the four-variable framework, this chapter's findings support the arguments of Weeks (2008) and especially Weiss (2013) about authoritarian states using domestic audiences to gain leverage in international disputes. In summer 2012, Japanese actions prompted Chinese leaders' need to show clear resolve to prevent Japan from making further moves to control the islands. Weiss argues that to be effective at signaling resolve to the foreign adversary, nationalist demonstrations must appear genuine and not government orchestrated or of the "rent a crowd" (p. 3) variety. On this score, the real-world protests were likely large and noisy enough to appear at least partially rooted in genuine popular grievance, and *Weibo* commentary certainly reinforced this impression. Yet while I agree with Weiss that part of Chinese leaders' motivation for allowing anti-Beijing *Weibo* comments, calls to action, and other strident remarks to spread certainly could have been rooted in their desire to gain international leverage, I differ in regarding such a motive as leaders' only or even primary motivation. If international leverage was the main goal, then (very real worries about protests getting out of control notwithstanding) why not lower censorship of certain comments in September as well as August?

Instead, the data speak to the primary importance of the *domestic* audience, with leaders seeking to placate angry hard-line netizen critiques in August, while seeing the opportunity (and the necessity) of shutting them down later in September.

Lastly, throughout this dissertation I have excluded leaders' possible ideational motivations for relaxing censorship from the analysis. Yet more than any other, the Diaoyu case suggests that the four-variable framework's instrumental logic, while valuable, may not completely capture elite thinking. In the *China Quarterly* article upon which this chapter is based, Allen Carlson and I note that Chinese leaders "lived within a social milieu (of their own making) in which negative perceptions of Japan had become naturalized" and that leaders' instrumental calculations "emerged from a cauldron of xenophobia" (p. 40). The key point is that instrumental, and ideational explanations for leader behavior under high-risk strategic circumstances are often not mutually exclusive. Instrumental rationality, in this dissertation centered around leaders' assessment of the material risks and benefit they face in imposing censorship or not – "material" in the sense of decisions having real-world consequences for leaders' survival – is a powerful mode of explanation. But subject to the survival constraint, this logic does not totally preclude more affective goals leaders might have such as the positive feeling they might obtain from allowing their own citizens to join the chorus of opposition to Japan. As mentioned in Chapter One, instrumental and ideational explanations are really at odds only when the latter make predictions about dependent variable outcomes that clearly differ from those of rational models, which does not occur in this chapter.

Chapter Six represents the last of three cases in which I have endeavored to show selective censorship in action across a range of issues, external circumstances,

and specific online sentiments. Next, the concluding chapter revisits Chapter Two's theoretical claims in light of the empirical evidence, considers alternative explanations, and considers the framework's applicability in contexts beyond *Weibo*, the year 2012, and China.

CHAPTER 7

CONCLUSION

7.1 Introduction

When and why do autocrats censor the Internet, especially social media? Just as importantly, when and why do they not? How are top leaders able to swiftly and decisively implement censorship interventions, coordinating orders across a plethora of bureaucratic and private sector agents? This dissertation has aimed to provide new insight into these questions via a study of China as the world's most sophisticated censoring regime. I began with Chapter Two, which considered the *why* of a varied and nuanced (rather than monolithic and crude) online censorship strategy. Contrasting with numerous formal and non-formal accounts of autocratic incentives to censor media or not, the chapter advanced a novel explanation for non-censorship: leaders temporarily relax control to signal responsiveness to acute popular demands that they themselves accept as legitimate. In an environment where the public expects tight social media control to be the norm, relaxing control can serve as a means of communicating to the public that leaders "get" how bad some problem is, and that they are on the hook to take action. In other words, the transparency of non-censorship itself shows state responsiveness to society. Yet leaders must balance such a motivation against other incentives such as the risk that an open Internet will help protesters coordinate, and thus *responsiveness benefit* is only one part of a four-variable framework that seeks to capture the nuanced and conflicting factors that affect the relative tightness of information control over time.

Next, Chapter Three, while not directly confirming or refuting the idea of *responsiveness benefit*, presented evidence that around 2011, top leaders came to view a more fine-grained (if in some respects, firmer) approach to managing social media as the right one. Although interviews with Chinese media professionals could not conclusively confirm any particular logic of control, they did establish that elites became increasingly strategic and unified in their approach to regulating *Weibo* and similar platforms. The rest of Chapter Three then focused on the *how* of Internet regulation by mapping, for the first time, the bureaucratic lines of authority that formed subsequent to the creation of the Cyberspace Administration of China (CAC) in 2014 as an offshoot of the State Council Information Office (SCIO). This analysis concluded that the centralization of regulatory authority over Internet companies in Beijing Municipality (under whose jurisdiction most companies operate), a generally symbiotic relation between companies and the state, and the transfer of oversight authority away from 'traditional' agencies (the police and the propaganda department) into the newly-enhanced CAC *Xitong* with a direct report to President Xi were all key factors in linking day-to-day censorship implementation with direction from the top. While treating the Chinese state as 'unitary' for purposes of *any* policy area is normally problematic, this finding gives warrant to think that on Internet regulation it is unified enough to make investigating whether a top-down logic of censorship exists a worthwhile endeavor.

After articulating and providing initial warrant for a logic of selective censorship, the dissertation then turned to three empirical analyses of online breaking incidents on Sina *Weibo* in Chapters 4-6 to illustrate Chapter Two's theoretical framework in action. Specifically, in each chapter/case I derived predictions of high, medium or low censorship disaggregated by individual issue frames or sentiments within each incident, and at different moments in time. Derivative of this,

I then predicted the dynamic relation between each sentiment category and the daily fluctuating censorship rate of relevant *Weibo* posts, as coded by what I assessed to be each sentiment's *responsiveness benefit* and *image harm* (combined into *credibility payoff*), as well as the incident-level *collective action risk* and *visible censorship cost* at different points during each incident. As both Chapters One and Two acknowledged, making predictions about how China's leaders are likely to view the benefit versus risk of censoring select sentiments – like Chapter Four's *Domestic vis-à-vis Foreign* or Chapter Five's *Critics* categories – is a demanding exercise that ultimately relies on the analyst's subjective judgment. That said, it is not as *ad hoc* as it might appear, since research in Chinese politics has done much to illuminate for what issues and under what circumstances leaders feel threatened (*image harm* and *collective action risk*), and when they feel the need to be transparent (*responsiveness benefit* and *visible censorship cost*).

While Chapters 4-6 do not (indeed, due to the research design and data limitations, cannot) prove that leaders acted on the basis of Chapter Two's four-variable framework, indirect evidence abounds. In Chapter Four's study of online discussion over the Ministry of Environmental Protection's reaction to the Beijing U.S. Embassy's release of its own air pollution statistics, comments that expressed the matter of reliable pollution data in more objectively scientific language were censored less than comments that used the dispute to criticize the Chinese government or compare it unfavorably to the U.S. This finding provides a prime example of *responsiveness benefit* and *image harm* in action, as the censors shifted over the course of the year from an indiscriminate approach to deleting related posts, to a more nuanced one of permitting commentary related to the PM 2.5 measurement standard while more aggressively suppressing commentary that directed attention toward central government responsibility. The fact that the Cyberspace Admin-

istration permitted the *Under the Dome* documentary to air for almost a week in 2015 despite its political sensitivity further supports the idea that at least on air pollution, officials recognize that the public's demand for pollution information and remediation is legitimate and urgent.

Within Chapters Four and Five, three of the framework's four variables factor in to some extent (with *collective action risk* the exception since it is low and invariant in these two cases). Yet each chapter emphasizes a different factor as primary. While Chapter Four's focus is on *responsiveness benefit*, Chapter Five – though definitely not neglecting both *responsiveness benefit* and *image harm* – especially highlights the role of *visible censorship cost*. Periodic revelations of shocking news at various points during the Bo Xilai scandal constrained top leaders' options in whether and how much to suppress online discussion. This is not to say that leaders had no control over whether news broke in the first place – in later instances, they themselves set the agenda by choosing when to go public with accusations against Bo – but once news did become public, *visible censorship cost* made rapid suppression of related commentary more difficult. The Bo scandal was China's highest-profile political scandal to date and the first of its kind during the social media era. The issue of elite-level corruption strongly resonated with politically aware Chinese citizens after years of reports of official extravagance and graft. Although leaders and the Sina Corp officials that answered to them had the ability to rapidly terminate discussion and commentary related to each breaking incident, doing so might have led online citizens to infer that corruption (and even sinister elite-level power plays) were more widespread and insidious than even the scandal's own events suggested. This was especially the case in the initial days after Wang Lijun's flight to the U.S. Consulate in February 2012, an event whose timing, at least, likely took officials in Beijing by surprise, and for which they

were probably ill-equipped initially to handle the leak of related news to the press. This case illustrates the crucial point that if highly visible, censorship itself can be embarrassing to the regime.

Finally, Chapter Six's study of the 2012 Diaoyu/Senkaku islands crisis entails aspects of all four factors but clearly stresses *collective action risk*, as it was the only case among the three where collective action on the ground was a serious possibility (and did in fact take place). The *Weibo* data clearly distinguish between lowered censorship in response to an initial surge of nationalist commentary in August – when many Chinese commenters criticized their own government for not taking a hard-enough line against Japan – and September, when street protests had swelled to hundreds of thousands of individuals and officials in Beijing themselves felt compelled to issue tough statements about Japan's claim to the islands. In the former instance, officials lowered censorship despite potentially high *image harm* to their perceived credibility to defend China's interests and honor, a factor potentially offset, however, by the *responsiveness benefit* they would derive from allowing even such hard-line comments some room. By allowing such raucous criticism, Chinese leaders showed just how seriously they took the charge that they were weak or feckless in defending the nation's territorial interests, and that they viewed such discontent as understandable. Nonetheless, the fact that real-world *collective action risk* was rapidly increasing (with demonstrations outside the Japanese Embassy in Beijing as early as August 15), likely militated against allowing such reduced censorship for long. And in September, the very real presence of collective action was an overriding factor that led censors not only to suppress government criticism, but even to delete calls for protesters to moderate their behavior where these calls referenced in-the-street activities like rioting, since such posts spread information about the collective event and ran the risk of encouraging

copycat behavior.

Taken together, Chapters 4-6 show patterns of variation in censorship and in sentiment categories associated with the four-variable framework that are consistent with its theoretical predictions. However, such evidence is correlative, and might be linked by a causal explanation different from the one articulated in Chapter Two. The next section revisits the alternative explanations I first raised in Chapter One, in light of the empirical chapters.

7.2 Internal Validity: Alternative Explanations for the Censorship Pattern

In this section I consider three challenges to the research design's internal validity in inferring a relationship between the four-variable framework, sentiment categories as proxies for these variables, and censorship: measurement issues; idiosyncratic event-specific explanations for the censorship pattern; and alternative logics of censorship.

7.2.1 Potential Measurement Issues: "Hidden" Censorship Not Observed, and *ReadMe* Error

Chapters 4-6 rely on a particular measure of censorship: the percentage of deleted *Weibo* posts in the dataset out of total topic-relevant posts, per day. Although post deletion is certainly an important aspect of the broader concept of

'censorship' or information control on social media, it is certainly not the only one. In fact, users' intended writing or image posting may encounter censor intervention under state direction at multiple points during the process. Ng (2014) observes several "pathways" that a post can take from the moment it is typed by the user and the "Submit" button is clicked, to the post's ultimate fate. These include: a) pre-publication censorship visible to the user, who receives a warning message saying that the content cannot be submitted because it contains one or more banned keywords; b) pre-publication censorship invisible to the user, who with no error message or explanation given, simply finds that he/she is unable to submit the post; c) embargoing of sensitive posts, where the content is allowed to be submitted but the user then receives a notice that it is subject to review by a censor, who can decide to let it appear publicly on the user timeline or to block it permanently; d) hidden embargoing, the same as c) except the user, while allowed to successfully submit the post, is unaware that it has been prevented from appearing publicly; and e) none of the above, where the user is allowed to submit a post, it immediately goes public, and then the user is later notified (or not!) that the post has been taken down, or is notified that his/her account has been deleted. Of course, there is also the sixth case: f) where nothing occurs and the post continually remains live as it would on a non-censored microblog site.

Of all these possibilities, the *WeiboScope* data only captures e), f) and c)/d) only if these latter two possibilities ultimately advance to stage e) or f).[1] Thus, the dataset is missing instances of pre-publication censorship. The consequence of the *WeiboScope* data not capturing these other forms of censorship is thus to underestimate the true extent to which users' intended comments are thwarted from becoming publicly visible over the long term. This is admittedly an issue if

[1]See Ng (2014), Figure 1 and Also King, Pan and Roberts (2014), Figure 1 for a graphical description of the above.

one's goal is to estimate what I call *total censorship*: the hypothetical fraction of *intended* posts (posts that a user wishes to write, whether allowed to or not) that are censored in some form divided by total intended posts. Because the *WeiboScope* data misses all intended posts that do not at least survive to the point of being publicly visible for a short time, this means that censorship estimates that rely on scraping only publicly visible posts – i.e. all non-experimental studies to date – will be biased downward.

This effectively renders difficult to impossible all observational designs by outsiders to accurately measure total censorship, since Sina and other Chinese Internet companies closely guard their internal (pre-publication) censored data. That said, to what extent does this issue threaten the dissertation's findings? I argue that the impact is less than might appear because the two most novel theoretical contributions – the existence of *responsiveness benefit* and *visible censorship cost* – both point in the direction of *reduced* censorship. If some amount of especially sensitive posts were in fact being pre-filtered along pathways a)-d) above in the incidents/issues of this dissertation, this would mean the true censorship rate was higher than measured in Chapters 4-6. However, as long as the censors are not suddenly substituting pre- for post-publication censorship, decreases in the latter also imply a decrease in the "true" rate, as Equation 7.1 suggests:

$$R^*_{total} = \frac{C_{pre} + C_{post}}{C_{pre} + C_{post} + P} \tag{7.1}$$

Where R^*_{total} is the total censorship rate of all *intended* posts, P is intended posts that are not censored at any stage (all of which appear in the dataset), C_{pre} is intended posts subject to pre-publication censorship, and C_{post} is post-publication censored posts, some fraction of which survive long enough to be captured by

WeiboScope (see Appendix B). As long as C_{pre} remains constant (or decreases) during a decrease in C_{post}, R^*_{total} will decrease. So what is required is that C_{pre} and C_{post} not be inversely related, where less post-publication censorship implies more pre-publication control for a given sentiment/topic and moment in time. The concern is not the gradual switchover from post- to pre-censorship over time or across different issues, but a rapid change within a single issue/incident and within a day or two. However, if this were happening, for sensitive incidents where we know censorship is widespread, we would also expect to observe a sharp drop in *observed* censored posts. Some such drops do occur in Chapters 4-6, but in all or nearly all cases can be accounted for by decreases in real-world news about the topic, or a more gradual natural decay in public interest toward the incident as users simply shift their attention elsewhere.[2] These drops are also usually accompanied by *increases* in observed total posts ($C_{post} + P$), which should not occur if C_{pre} is increasing and there is no other external shock that triggers a surge in overall post volume. This means that in all likelihood, the inferences I made in Chapters 4-6 regarding the censorship rate and its relation to various sentiment categories continue to be valid.[3]

A second measurement concern has to do with the *ReadMe* estimates. The specific worry is that measurement error in some of *ReadMe's* estimated category proportions might be correlated with the measure of censorship; conversely, if *ReadMe* error is random or haphazard, this will increase coefficient standard errors in regressions with censorship as dependent variable, but will not bias point estimates, and any significant results that obtain despite the large standard errors can be taken seriously. I test this possibility by regressing the censorship rate on

[2] As King, Pan and Roberts (2013) note, social media discussion of most viral topics tends to be "bursty" and short-lived even in the absence of censorship.

[3] Subject to the assumptions about the true rate of *post*-publication censorship as measured by the *WeiboScope* data, which are examined in Appendix B.

the *ReadMe* error when compared with the baseline of human-coded category proportion estimates. Across the coding exercises from Chapters 4-6 there are a total of ten discrete dates or date ranges for which both hand-coded and *ReadMe* estimates are available – with multiple sub-estimates by sentiment category available for each. I pool the estimates across j sentiment categories to derive the average *ReadMe* error for each date or date range t:

$$AvgError_t = (1/n)\sum_{j=1}^{n} ReadMe_{j(t)} - (1/n)\sum_{j=1}^{n} HandCoded_{j(t)} \qquad (7.2)$$

This yields ten estimated average errors (by date or date range). I then do a simple bivariate regression of the ten corresponding censorship rates on these errors. Results are highly insignificant ($\beta = -.01, p = .904, N = 10$). Using absolute-valued rather than signed errors yields similar results, as does randomly sub-sampling sentiment categories, computing their measurement errors, and regressing the corresponding dates' censorship rates on them. Based on this exercise, there is no evidence of any correlation between *ReadMe* measurement error and the censorship rate.

A final concern has to do with the (post-publication) estimated censorship rate itself. Appendix B lays out the assumptions necessary for the correction formula I use to account for the *WeiboScope* data's under-estimation of censorship. The key assumption is that the behavior of censorship (especially its speed after a post first goes up, and at different times throughout the day, e.g. morning, evening etc.) is similar enough in *WeiboScope* to the data used by Zhu *et Al* (2013). Both datasets draw a sample of high-profile *Weibo* users in 2012, and both capture many of the same breaking incidents that occurred during that year. As a first pass, it is reasonable to believe that since the same team of trained censors for Sina

were active under defined policies and procedures during 2012, censorship behavior should be comparable between datasets. However, to go beyond this I carry out a series of "face validity" checks for how we would expect the *WeiboScope* measure to behave (irrespective of specific case or issue) if it is indeed capturing real-world censorship; for example, to increase during real-world collective action events but not, say, in the middle of the night when *Weibo* posters and censors alike are likely to be less active. These are reported in Appendix B.4.

7.2.2 Idiosyncratic Case-specific Explanations

The factors in the four-variable framework are all based on varying degrees of theoretical and empirical support from previous work. That said, they are all still in a theory-building rather than theory-testing stage. Since Chapters 4-6 are then intended to illustrate rather than test the salience of these factors, and although the chapters each show clear variation in censorship that is consistent with the framework's variables, the door remains open to at least three alternative explanations. The first, broadly speaking, entails such possibilities as bureaucratic incompetence, bureaucratic or Internet company interests below the central level, or the corruption of censors – all micro-level, non-systematic explanations. Chapter Three showed that such explanations, while still possible, became increasingly unlikely subsequent to reforms launched in 2011 and are especially improbable for high-urgency incidents like those in this dissertation. But what if the censors and intermediate officials *were* unified under a decisive chain of command and were swiftly acting in each case according to clear orders, but the orders were *sui generis* to each incident and did not reflect *any* logic applicable across incidents or over time, including the four-variable framework's? What if leaders had completely

different ideas for how to manage censorship during the Beijing U.S. Embassy air pollution dispute, the Bo Xilai scandal, and the Diaoyu/Senkaku islands protests? After all, such a starting point would certainly be the default for scholars of Chinese environmental politics, elite/factional politics, and popular nationalism. Indeed, many of my interviewees resisted the notion that these three cases (and others) were comparable. Discussion of air pollution was dismissed as simply 'not that sensitive';[4] allowing frank discussion of the Bo scandal in its early weeks was mere factional politics or top leaders' desire to humiliate Bo (rather than signaling to the public where they stood in any broader sense);[5] and allowing government criticism during the 2012 Diaoyu protests was a one-time, unique case where CCP elites wanted to appear especially tough prior to the leadership transition, rather than a long-term strategic possibility irrespective of the specific historical juncture.[6]

Aside from sounding very *ad hoc* to a social scientist, these explanations run counter to accumulated evidence thus far that certain common factors across issues *do* underlie censorship decisions. For example, leaders view collective action as risky whether it concerns politically charged events like nationalist protests, or relatively innocuous activities like the public statements of dissident artist Ai Weiwei, who is not considered mainstream in China but still has the potential to mobilize some protest (King, Pan and Roberts 2013; 2014). Similarly, Esarey (2013) and Tai (2014) demonstrate by analyzing leaked propaganda directives that a wide range of topics embarrass or threaten leaders (i.e. cause *image harm*), thus de-emphasizing the uniqueness of certain topics over others. *Responsiveness benefit* and *visible censorship cost* have not been empirically tested as rigorously as the above two factors to date, but the above examples do suggest that the burden of

[4]Interview #14, BJ, 10/14/14.
[5]Interview #25, SH, 12/13/14.
[6]Interview #5, BJ, 9/18/14.

proof should lie with skeptics or issue experts to show that broader logic(s) that transcend individual issues do *not* exist, given that the Xi Administration views comprehensive Internet control as vital to CCP survival. That said, further research that includes more case studies across new issues and over time is certainly valuable in the effort to address critics' concerns.

7.2.3 Alternative Logics for Selective Censorship Revisited

In Chapter One, I raised two alternative logics other than *responsiveness benefit* (and *visible censorship cost*) for why top leaders might relax control over online censorship: the 'safety valve' metaphor (Hassid 2012); and various 'information-gathering' logics (see Lorentzen 2014). Hassid (2012) claims that social media serve as a 'safety valve' when print media bring a controversy to public awareness, allowing angry netizens to release frustration about the issue. During fieldwork, some interviewees did mention the Internet as 'safety valve' (Mandarin: *jian ya fa*).[7] And indeed the idea that the government was allowing netizens to vent frustration is at least plausible for all three cases in Chapters 4-6. However, once again the problem with the 'venting' explanation is that in the politically charged atmosphere of Chinese social media and especially *Weibo*, so-called 'venting' is never merely a series of isolated complaints, but a powerful common knowledge-generating phenomenon. One could argue that the state is aware of this and allows venting to occur briefly enough that longer-term public coalescence around some 'truth' – e.g. the state is to blame for some problem – does not really occur, but this assumes that leaders have near-perfect foresight to gauge how fast and under

[7]Interview #4, BJ, 9/16/14 provides the clearest example of an interviewee who believed the 'safety valve' hypothesis.

what circumstances the public will solidify its shared view.

In the volatile world of social media, this assumption is implausible. Put differently, social media discussion of political controversies is *risky* for an authoritarian state that relies on media control and propaganda as extensively as China does. Allowing citizens to vent among family and friends in offline space or in less 'public square' online applications like WeChat is far less risky than permitting *Weibo* discussion. Chapter Four aptly illustrates how mere 'venting' can quickly turn into collective demands. One could interpret increases in blogger speech about PM 2.5 and other health concerns as just that – public fretting over air pollution's harmful effects. But because the controversy was over the contrast between the U.S. Embassy's release of more valid monitoring statistics versus the Ministry of Environmental Protection's longstanding reluctance to do so, public discussion quickly gave rise to an expectation of increased government transparency and for official efforts to improve air quality. The other two cases similarly testify to the riskiness of *Weibo* discussion. In Chapter Five's discussion of the Bo scandal, it would be hard to maintain that allowing netizens to 'vent' about Bo's corruption was innocuous when a nontrivial percentage of commenters used the relative openness not to spew invective at Bo himself, but to use the incident to raise broader questions about the political system. And in Chapter Six's analysis of the Diaoyu protests, netizens' anger in mid-August against their own government for not taking aggressive enough actions against Japan could not be merely blowing off steam but rather was a genuinely dangerous cocktail, given how potent the accusation has been in previous historical periods that incumbent Chinese governments failed to stand up to foreign exploitation.

A second alternative explanation is actually a group of related theories about

the 'information-gathering' benefit autocrats are said to derive from allowing some media openness, of which Lorentzen's (2014) argument is a prime example with respect to China. In Chapter One I suggested that this explanation probably does hold considerable merit in accounting for less-than-total media control in China, including social media. The key point, though, is that the four-variable framework and an information-gathering logic are not at all mutually exclusive. Rather, each applies to different categories of online incidents. Instances of localized protest, such as labor strikes, village protests over government land seizures, and ethnic unrest in China's far-Western region constitute the vast majority of protest incidents in the country. Due to their number and local official incentives to withhold information, leaders in Beijing also have great difficulty keeping track of them. The sort of high-profile incidents in this dissertation, in contrast, all involve issues of which central leaders were likely well aware. Top officials, especially environment officials, almost certainly knew that the problem of air pollution was worse than the public was aware of prior to 2012. Central leaders may not have been aware of the exact timing of Wang Lijun's flight to the U.S. consulate in Chengdu, but long before the scandal became public they were able to monitor Bo and Wang via internal channels. And Beijing officials both knew that activists were setting sail for the Diaoyu islands as early as August 12, 2012 (if not before), and in all likelihood knew (from past experience) that an activist landing on the islands would provoke both a Japanese intervention (in this case, detaining them), and a strong public reaction once the landing and detention became public.

Assuming that leaders and the Internet bureaucrats under them did understand the nature of public discontent in each of the above issues, why allow any uncensored discussion to take place? The argument hinges on the specific 'information' officials would hope to glean by opening the floodgates. One could argue

that officials sought to learn not the existence of public discontent, but its intensity. Yet Lorentzen's (2014) model and related models (Little 2016; Shadmehr and Bernhardt 2015) tend to treat censorship as a binary choice – either some news or protest event is suppressed, or not. They do not examine the multi-period effect of not censoring on citizen perceptions of the issue at stake and of the regime: in other words, a 'snowball effect' where the risk to the regime grows with each hour or day that content is left uncensored. Because spikes in political discussion are highly "bursty", however, leaders should be able to measure the intensity of public sentiment within the first few hours or day of an initial surge before choosing to shut it down. They should have no incentive (and strong disincentives) to allow discussion to continue and snowball over a longer period. Yet this is precisely what we observed in Chapters Four and Five. In the air pollution case, reposts of AQI monitoring data and discussion of PM 2.5 went relatively uncensored for months during the first half of 2012 as the crisis escalated toward its June peak. And in the initial weeks of the Bo scandal following Wang Lijun's flight, censorship remained relatively low by the standards of sensitive incidents ($< 40\%$) even as skepticism about Wang's connection to Bo and Bo's misdeeds proliferated. If leaders only wished to gauge the intensity of public discontent in these two cases, they could have done so by lowering censorship for a much briefer period than actually occurred.[8]

What About the Big V? An Individual-level Alternative Explanation

Finally, this dissertation has not gone into great depth regarding the back-

[8]Chapter Six does appear consistent with an "intensity of discontent" story, as the sharp reduction in censorship in August only lasted a few days and was only very low for one day. But this could be explained by the high risk of real-world collective action for nationalist protests, relative to the other cases.

grounds, personalities and followings of the many celebrity Big V that form the *WeiboScope* sample and that played a major role in facilitating information flows during the three incidents in Chapters 4-6. As the Big V were key to *Weibo's* liveliness in 2012, one could argue that any account of censorship variation that does not account for who they are, who follows them and how they differ politically is incomplete. Indeed, it could be argued that an in-depth study of individual Big V – whether through in-person interviews, or simply by reading their personal *Weibo* feeds over a long period – would have served three purposes. First, it would explain why certain Big V were so willing to criticize the government or top officials, thus playing a major role in generating *responsiveness benefit*. Second, an individual-level study of the Big V could help explain why *Weibo* was in 2012 (and perhaps still is) so vibrant an engine of common knowledge generation and viral information. Third, focusing on the Big V might have enriched contextual knowledge of individual cases, providing additional insight into both the sentiments that prevailed in each, and their censorship pattern.

While a study of the Big V toward the three purposes above certainly could have provided additional context for the dissertation's argument, doing so was not essential. There are two main reasons why this is the case. First, although in going through Chapters 4-6 readers (and this author) cannot see the identities (except for a brief mention of real estate developer Pan Shiyi in Chapter Four) of the most prominent Big V who wrote and re-posted comments, the key point is that both other microbloggers and non-celebrity *Weibo* users, and Internet company and government censors would have been well acquainted with these individuals and would have been able to observe who said what in real time, and to react accordingly. If one or a handful of super-influential Big V were responsible for singlehandedly driving entire sentiment categories, this would be an issue. How-

ever, the incidents in Chapters 4-6 all involved a plethora of voices – a fact I could ascertain by looking at the anonymized (but consistent) User ID number attached to each post. Thus, rather than choosing to focus on and censor individual Big V alone (which could have occurred in some cases but would have had limited effectiveness for stopping entire topics), the censors would have had to target the collective spread of such topics and sentiments regardless of originator.

In short, censors likely were primarily reacting to the message rather than the individual. Yet even if censors were to target particular individuals, they would likely do so as a shortcut method for finding sensitive content (since some bloggers have a reputation for pushing the limit) rather than based on who the bloggers were.[9] Of course, who the Big V are and what they say are closely intertwined, but as long as they serve as consistent voices for particular views (in response to breaking news events), we need not dive into each Big V's background and professional identity to understand how they are censored. Indeed, conceptualizing the issue as one of including appropriate "control variables" in the censorship model, the Big V's "individual identity" variable was (relatively) constant over the short period of time (the year 2012) in which the *WeiboScope* data were generated.

Because the identity variable is constant and both other bloggers and the censors observe and account for it as a constant, it cannot explain fluctuations in censorship across incidents, sentiments, or 2012. "Measuring" the Big V's identities and which Big V spoke the loudest at certain times would therefore not be measuring an "omitted variable" key for modeling censorship, but rather constitute an entirely different project. That said, although not essential for this project's

[9]For example, during Chapter Four's analysis, my co-author and I noticed that Pan Shiyi was censored to varying degrees at different points during the year, where the difference was not changes in Pan's overall public identity, but in what he said – early in 2012 his comments were rather mild, but became increasingly provocative around the June peak of the dispute, which prompted the censors to react.

specific goal of elaborating a theoretical framework of *macro*-level influences on social media censorship decision-making, inquiry into the Big V as individuals and as an activist social group is important to address larger questions about social media's longer-term impact on Chinese politics and state-society relations. Fu and Chau (2013) find that a small subset of all microbloggers (4.8%) generate over 80% of original posts, and that "volume of followers is a key determinant of... reposting messages, being reposted, and receiving comments" (p. 1). With so much online influence concentrated in so few hands, future work involving detailed reading and observation of China's leading microbloggers is vital to assessing *Weibo's* and other platforms' staying power as conduit for popular discontent, particularly given the recent crackdown on the Big V under Xi, which the next section discusses.

7.3 External Validity: Beyond *Weibo*, 2012, and China

This dissertation has set deliberately narrow scope conditions for studying censorship, focusing on one social media platform, during one year, in one country. The advantage of this move has been to allow for carefully controlled comparison between the cases in Chapters 4-6, and to pinpoint top leaders' intentions and their bureaucratic capacity to implement information control at a discrete time point in Chapter Three. Because the *Weibo* platform's specific characteristics, the macro-political circumstances of 2012 and China's world-leading status as censoring regime all *do* matter in explaining censorship variation, relaxing these constraints during analysis would have yielded an indeterminate research design. By holding them constant, I was able to focus on answering research questions two through four from Chapter One: why Chinese leaders came to view more nu-

anced censorship capabilities as desirable around 2011-12; how they acquired the capabilities to implement a fairly unified censorship policy across numerous intermediaries; and showing variation in individual cases consistent with a selective censorship logic. If we relax these constraints, how well would we expect the empirical findings to translate to other social media services, subsequent time periods, and other countries? The following sections consider each of these possibilities.

7.3.1 Beyond *Weibo*: Other Chinese Social Media Platforms

Chapter One posed the question "why social media?" to narrow the focus of inquiry from the entire Internet, to a vague interrelated set of communication technologies that allow the rapid, networked exchange of information among 'friends' or 'followers', often through a 'news feed'-like mechanism of continuously updated content. Microblogs such as Twitter and *Weibo* are the exemplar of such communication as it pertains to public political discourse. Sina *Weibo* is not representative of all Chinese microblogs but rather was chosen because nowhere else on the Chinese Internet were the trade-offs that rulers faced between social media's benefits and risks as stark as they were on this platform in 2012. As Chapter One discussed, *Weibo* is this politically salient due both to its viral technological characteristics, and to the different actors that inhabit it: ordinary users, Sina Corporation, the 'Big V', and journalists. What about other Chinese platforms that have similarly viral technology for re-posting content, and similar constellations of participants? An initial answer would be to say that no such platform exists: in Chinese Internet history, only *Weibo* has ever brought together the right technology and participants

to serve as a true 'public square' for political discussion.

This is obviously not a very satisfying answer. While no Chinese domestic platform in recent memory has approached *Weibo's* political relevance in 2011-12, theoretically speaking for any service to be 'similar enough' for the four-variable framework to be applicable, it must be capable of generating nonzero *responsiveness benefit*, *image harm*, *collective action risk* and *visible censorship cost* for the central government. Equivalently, it must a) technologically enable the rapid and widespread (national) generation of common knowledge, and b) be populated by individuals with the motivation and ability to generate that knowledge (such as celebrity bloggers and journalists). Other microblogs present in China in 2012 may have realized a), but fell short on b). Tencent's *Weibo* service, which grew as an offshoot of the popular (in the 2000s) instant messaging service QQ, had much of Sina *Weibo's* early functionality but was more centered around social relationships. For a time it appeared like Tencent might compete directly with Sina's service, but it never matched Sina in terms of monthly active users: 2012 was the closest it came with 277 million active users to Sina's 287 million.[10] However, Tencent *Weibo's* followers were concentrated in poorer inland cities rather than wealthier coastal ones, and the platform initially attracted less attention from 'Big V' as a forum for lively debate.

Due to Tencent *Weibo's* lack of elite-city "buzz" as well as Tencent's difficulties with monetizing the platform, it ultimately closed its *Weibo* business division in 2014, perhaps in part because of the surging popularity of the company's other product WeChat (*Weixin*), which I discuss at length below. First, though, one

[10]Source: Tech in Asia. https://www.techinasia.com/tencent-weibo-registered-users-540-million. Although "monthly active users" is the gold-standard metric of social media site popularity worldwide, comparing statistics across platforms is fraught with difficulties such as different definitions of 'active', and how companies deal (or not) with number inflation due to 'spam' or 'bot' accounts.

other platform bears mentioning: RenRen (Mandarin for "everyone"), which unlike Tencent *Weibo* is still being actively developed by its parent company, Renren, inc. RenRen is not a microblog but rather a social networking service similar to Facebook – in fact its original name was *Xiaonei* (meaning "within the schoolyard"), an emphasis that paralleled Facebook's early growth and expansion on college campuses. Despite its early popularity (peaking at just under 60 million monthly active users in 2013), it has since seen a slow decline and has never matched the user base or popularity of (Sina) *Weibo* or WeChat.[11] Somewhat unlike Facebook, RenRen never successfully diversified away from its focus on immediate friendship networks to become a more blog-like platform favoring celebrities and other prominent voices.

As platforms centered around friendship ties rather than bloggers and followers, it is debatable whether RenRen and similar platforms allow for common knowledge generation as effectively as microblogs. There has been some empirical support for the role of Facebook pages in coordinating protests during the Arab Spring (Hussain and Howard 2013), but coordinating protest logistics in the (relatively) un-censored Arab context is very different from fostering sustained political engagement in China and in situations that do not entail massive collective action. This then prompts consideration of the only social networking platform in China that has matched (and indeed exceeded) Sina *Weibo's* popularity: WeChat, which had an astounding 806 million active users in 2016, the overwhelming majority of which were in mainland China.[12]

Does WeChat have the potential to serve as China's next big political 'public

[11]Source: Tech in Asia. https://www.techinasia.com/china-facebook-social-network-renren-losing-users-fast

[12]Source: Tech in Asia. https://www.techinasia.com/wechat-and-the-bamboo-ceiling. Even if this figure is inflated, it is still clearly the predominant social platform of any kind in China.

square'? Since WeChat has surged in popularity entirely during the post-2012 Xi Jinping era, in reality such a question cannot be walled off from the following section's broader consideration of Internet control under Xi. For argument's sake, though, let us consider the counterfactual of WeChat, which was only launched in 2011, being as popular in 2012 as it is currently (early 2017 at the time of writing). Could WeChat discussion have produced the same costs and benefits – defined by the four-variable framework – in each of the dissertation's three cases, assuming that the "on the ground" facts of each incident were the same? The answer is probably not, if for no other reason than because of certain inherent features in the software that make the viral spread of information much more dependent on strong personal ties and more similar to real-world social interaction. To begin with, WeChat users cannot "search" for other individuals to connect with by keywords; one must know the exact user ID of the friend's WeChat account, or scan her/his QR code.[13] This makes it difficult to follow celebrity users, especially since the user receiving the request must choose to accept the new individual as contact – this is a 'reciprocal' social network, where ties only exist if both parties agree.[14]

A second type of functionality on WeChat is called "Public Accounts" of which there are two main sub-types: subscription accounts, and service accounts. While the exact features of each are nuanced and not worth describing in depth here, both have major limitations for those wishing to virally spread content. Both account types are able to accumulate "followers" and are findable through the WeChat main search function, though only if the user has a highly relevant keyword

[13]A QR code is a two-dimensional form of barcode now frequently used on mobile phones for various purposes, including product discounts, sporting event tickets, and to provide other individuals a quick and easy means to 'friend' one's social media account.

[14]The no-searching restriction only applies to finding *individual* accounts and not "Public Accounts", as I discuss below.

in mind (Tencent has been careful to avoid irrelevant or 'spam' search results). Both individuals and media outlets can have subscription accounts while registered businesses or organizations in China can have service accounts. Administrators for subscription accounts can post an update (just as one would update on Facebook or Twitter) once per day, but the update will not be "pushed" to followers' main feeds (as is the case on *Weibo* and Twitter), requiring users to instead browse into the subscriptions menu to read the content, and in practice, reducing their responsiveness to the update. Conversely, service account administrators' updates are immediately visible in followers' feeds, but they are only able to post one update per week.

These two examples of specific WeChat functions illustrate just some of the technological limitations Tencent has built into their product that make user-generated content creation and dissemination a more networked, personal and "curated" (i.e. small volume and restricted) process than on microblogs. While Tencent has good commercial reasons to configure the platform this way, such as improving the user experience by preventing "spam" information from marketers and others, it also clearly has the effect of limiting how fast and how far political influencers are able to spread commentary about breaking news, except through one-to-one and small group chats.[15] In sum, because WeChat's inherent structure limits the rapid spread of common knowledge, my theoretical prior is that the four-variable framework should not apply, and selective censorship along these lines should not occur. In an initial study, Ng (2015) finds that censorship does occur on public accounts, but at a lower rate than on *Weibo* and with a greater

[15]During fieldwork, multiple interviewees suggested that the WeChat one-on-one and small-group "rumor mill" might ultimately generate greater underground and bottom-up political change in China than top-down broadcasts by 'Big V' had been able to achieve. This possibility certainly deserves further research, but requires different research methods (likely ethnography or participant observation rather than big data), and different theoretical underpinnings than found in this dissertation.

emphasis on "fake news" and "rumors". Further research is needed to qualitatively and quantitatively describe, and eventually theorize about the potentially different logic of censorship in WeChat space.

7.3.2 Beyond 2012 in China: The Xi Era

Since taking office in November, 2012, President Xi has placed considerable emphasis on tightening control over traditional and new media, relying on a combination of ideological, legal/regulatory and bureaucratic means. The centerpiece of this approach has been a much firmer ideological line regarding what Xi views as the proper role of digital media – companies, news professionals, and individual bloggers – in China's socialist system. Along lines reminiscent of the Mao era, a secret Party communiqué, the so-called "Document No. 9", identified the Internet as a major locus of ideological risk and "mistaken thinking" for the CCP.[16] Xi himself elaborated on this concern in an August, 2013 speech to the National Propaganda and Ideology Work Conference in which he referred to the Internet as "the main battlefield for the public opinion struggle."[17]

These documents and comments indicated that a major campaign for the Party to suppress dissenting and critical voices, especially on microblogs, was underway as of early 2013. Ren Xianliang, who was promoted from the Shaanxi Provincial Propaganda Department to become Vice-Director of the SIIO (later CAC) that year, was an especially outspoken proponent of silencing the Big V. Creemers

[16]Source: Translated by ChinaFile, available at http://www.chinafile.com/document-9-chinafile-translation.

[17]Source: China Copyright and Media. "Xi Jinping's 19 August Speech Revealed?" https://chinacopyrightandmedia.wordpress.com/2013/11/12/xi-jinpings-19-august-speech-revealed-translation/. Accessed 4/20/17.

(2017) notes that Ren viewed the Big V's influence and audience reach as having surpassed print media and that "[we should] warn those that should be warned, shut up those that should be shut up, and close those that should be closed" (Ren 2013). Yet officials' plans for re-gaining ideological turf in the blogosphere went far beyond silencing the Big V. Rather, the CCP has aimed to *replace* them with a loyal commentariat of bloggers that simultaneously a) do not cross political bottom lines, b) consistently spread "positive energy" – a euphemism under Xi for silencing critical or dissenting online views using positive-sounding pro-CCP rhetoric – and c) do so in an organic, engaging way that leverages bloggers' standing as public figures. For this effort, news organizations like *People's Daily* are cultivating so-called "Medium V" (Ke Li 2015), who are typically "professors, high-ranking editors and journalists, or lawyers and experts" (p. 19). These individuals are supposed to use their social credibility to organically promote pro-Party views, but without the egotism, critical tone and above all, political disloyalty said to characterize many of the Big V.

Xi's and his subordinates' attempt to field a "national Internet team" of commentators to guide microblog opinion in a pro-CCP direction is only one prong of a multi-pronged strategy of direct Party intervention in online space. Another, which Chapter 3 analyzed in depth, has been Xi's tendency to consolidate regulatory power in the hands of *Party* organs rather than governmental ones, with the Central Leading Group on Informatization and Internet Security a prime example. In another example, the CAC, though technically also a state agency (i.e. the SIIO), has been largely staffed with propaganda cadres, many from the Beijing Municipal Propaganda Department. This in turn has lent an ideological tone to efforts to contain and regulate dissent – the state seeks not just to proscribe destabilizing or anti-CCP speech, but to prescribe morality. Cui and Wu (2015)

find that state-run media editorials often justify Internet control measures in terms of promoting society's "moral goodness". What constitutes such goodness is, of course, solely defined by the CCP. This role for the central state as both guardian and promoter of public morality bears resemblance to the Mao era in that central propaganda authorities are not merely broadcasting approved moral/political content, but attempting to enlist lower-level cadres and ordinary citizens alike in actively socializing it, and indeed to transform those very citizens into committed foot soldiers (Yang 2014).

A final emphasis in the Xi era has been on prompting both Internet companies and users to exercise "self-discipline" not only in implementing (self-) censorship, but in proactively policing their own or the community's behavior. Internet companies are now strictly liable, for example, for the content of video uploads, and companies like Sina have finally been pressured into implementing real-name user registration, a move they long resisted (Creemers 2017). And in an effort to head off further regulations and sanctions, Sina has developed community behavior standards that warn users against violating the same vaguely-worded "bottom lines" as the CAC has emphasized.[18] And according to a ruling by the Supreme People's Court, Internet users can now be imprisoned for up to three years if a sensitive post they write is retweeted over 500 times or receives over 5000 total views.[19]

The end result of these ideological and policy shifts has been to greatly restrict

[18]Source: The Next Web. Jon Russell. "Sina Weibo to introduce 'user contract' on May 28 as China's microblog crackdown continues [Updated]" https://thenextweb.com/asia/2012/05/09/sina-weibo-to-introduce-user-contract-on-may-28-as-chinas-microblog-crackdown-continues/ Accessed 4/20/17.

[19]Source: Supreme People's Court. "*Guanyu banli liyong xinxi wangluo shishi feibang deng xingshi anjian shiyong falu ruogan wenti de jieshi*" (Interpretation Concerning Some Questions of Applicable Law When Handling Uses of Information Networks to Commit Defamation and Other Such Criminal Cases). September 6, 2013. https://chinacopyrightandmedia.wordpress.com/2013/09/06/interpretation-concerning-some-questions-of-applicable-law-when-handling-uses-of-information-networks-to-commit-defamation-and-other-such-criminal-cases.

the space available for genuine dissent or even criticism of official policies within "public" online spaces, especially Sina *Weibo*, while simultaneously co-opting these spaces to support the CCP's renewed emphasis on "guiding" online public opinion rather than merely censoring undesirable sentiments. While independent academic data to quantify the crackdown are lacking, a leading state media online research center, the People's Daily Public Opinion Monitoring Center, observed a 25% decline in posts in a sample of 100 Big V in September 2013, only one month after the arrest of prominent Big V and outspoken government critic Charles Xue on what were viewed as politically motivated charges of soliciting prostitution (Ke Li 2015). Yet even as the CCP under Xi has sharply limited microblogs as independent space for public deliberation, it has expanded efforts to co-opt public input through more institutional channels. For example, the Central Commission for Discipline Inspection, which is responsible for carrying out President Xi's sweeping anti-corruption campaign, launched a mobile app in which ordinary individuals could directly report instances of corruption.[20] Such institutionalized "public feedback" e-governance mechanisms are far more consistent with Xi's emphasis on the Party harnessing online space to cultivate citizen loyalty to and trust in the central state, as opposed to the more "chaotic" environment on microblogs the new leadership viewed as prevailing under the Hu-Wen administration.

In short, while during prior reform-era leadership transitions there was often reason for China scholars to assume more continuity than change in longstanding CCP policy priorities like media and propaganda, the above ideological and regulatory trends since 2013 are too marked to take continuity as a working assumption any longer. Given recent evidence, the burden of proof now falls on proponents of 'selective' or 'strategic' censorship to show that the Xi administra-

[20]Source: China Daily - U.S. Edition. July 21, 2015. "Mobile app joins toolbox in anti-corruption effort."

tion sees any substantial value – instrumental or otherwise – in allowing genuinely independent deliberation on *Weibo* and similar platforms. The question is not "what strategic logic" the current leadership is following as regards flexibly managing outbursts of public criticism, but rather whether they see any strategic value at all in allowing more of the sort of criticism that occurred in this project's three cases, even when not linked to collective action and when not directly criticizing top leaders themselves. In fact, a major priority for the CAC and other propaganda authorities in managing microblog opinion since 2013 has been channeling such "negative" displays of emotion-laden criticism during breaking incidents – e.g. government criticism during the 2011 Wenzhou train incident – in a more "positive" direction. Such an approach entails de-emphasizing *responsiveness* to public grievances in favor of the "national team" steering discussion in line with the Party's positive rhetoric, so as to silence or drive grievances away. Instead, citizens are encouraged to submit complaints or reports of government misbehavior via the above-mentioned apps and portals, or as a comment (not broadcasted post) on *Weibo* or WeChat.

What are the implications of these trends for the dissertation's theoretical framework? An answer can be sought on two levels: theoretical, and empirical. On a theoretical level, it is precisely *Weibo's* role as common-knowledge generating platform that enables a tolerant central leadership to derive *responsiveness benefit* from relaxing control. If the Xi administration does not view sometimes allowing a certain degree of deliberative space online as legitimacy-enhancing – as some scholars have argued was the Hu-Wen leadership's approach (see He and Warren 2011; Lewis 2013) – then it will not factor *responsiveness benefit* into its censorship decision-making as the framework would predict. This does not mean that Xi and the propaganda officials under him do not view public input and even the expres-

sion of "public opinion" (in CCP co-opted channels) as useful; on the contrary, the CCDI's claim to have relied on numerous online tips to launch anti-corruption investigations, as well as the proliferation of official *Weibo* and WeChat accounts suggest that the leadership increasingly values controlled mass input. The question is instead whether, in addition to these venues, Xi and his people view any legitimacy-enhancing benefit to lively discussion in the digital public square.

Given the ideological and regulatory tightening since 2013, if it still exists such space is likely to be more limited than before regarding the issues and moments where "negative" or critical sentiments are tolerated. Yet I argue that occasional openness is unlikely to vanish in all instances. Returning to the four-variable framework, *responsiveness benefit* may be void in the Xi era, because it relies on ruling elite views of the merits of limited liberal discourse. But there is no reason that *visible censorship cost* – the other factor in favor of looser censorship – should not continue to apply, especially during highly visible crises and disasters. Major public health and safety crises in the 2000s, notably the 2003 SARS epidemic and 2008 Wenchuan Earthquake, impressed upon propaganda officials the need for some transparency during acute crises, both for reasons of public health and safety and to avoid the spread of rumors about the actual situation or about the government's responsibility (Chen Lidan 2008; Chen Ni 2009). A more recent example was the massive chemical explosion in Tianjin in 2015, in which some degree of non-Xinhua reporting (and of course, posting and reposting of news content on *Weibo*) was tolerated.[21] In the Tianjin case, whether by accident or design the censors seem to have shown some initial tolerance of discussion in the

[21]See leaked guidelines issued by the SIIO, Tianjin Municipal Propaganda Department, and an unidentified local propaganda authority. The guidelines allowed websites to re-post "authoritative sources" rather than restricting them to Xinhua copy as is typical in instances of tight control. Posted by China Digital Times. August 13, 2015. http://chinadigitaltimes.net/2015/08/minitrue-explosions-in-tanggu-open-economic-zone-tianjin/. Accessed 4/20/17.

explosion's immediate aftermath.[22]

Beyond illustrating the potential uniqueness of large-scale crisis situations in how the CCP handles outbursts of criticism, the Tianjin case also poses an empirical question: how many other episodes of partially relaxed control (if any) have occurred on *Weibo* since 2013, and if many exist, do they challenge the conventional wisdom that *Weibo* as political public square is "dead"? Second, within such instances of lowered censorship, who has replaced the Big V as re-broadcasters of news and commentary? Although the Big V were an easy target in the 2013 crackdown, data collection and analysis efforts should focus on the "Medium V" as well as commercialized media outlets as alternative key agents in spreading critical information. To be clear, my expectation is that future *Weibo* data collection of breaking incidents that occur during Xi's remaining years in office will indeed reveal greatly reduced "public square" space for anyone, Big V or not. But despite considerable evidence of a Party-centric approach under Xi to assert online ideological and technological control, I do not think that Chapter Two's theoretical framework has become entirely irrelevant. Just as in the late Hu years, Internet and propaganda cadres will continue to be sensitive to *collective action risk* and *image harm*. And if contrary to my expectation, *visible censorship cost* is no longer as motivating a factor to avoid heavy-handed censorship, this will prompt substantial inquiry as to why. One reason it could matter less is if the "national team" of cadres and loyal social media commenters becomes so skillful at steering discussion in a pro-Party direction, even during major crises, that the physical deletion or blocking of dissenting posts is no longer as pressing. This would resonate with Chen and Xu's (2016) formal model result that "[allowing] public communication... disorganizes the citizens or strengthens their disagreement if, through

[22]Source: The Straits Times (Singapore). August 21, 2015. "Social media abuzz as netizens poke and prod."

communication, they find themselves split over government policies" (p. 1). If online observers think the "national team" voices supporting the CCP's preferred interpretation of a crisis are genuine and those voices appear to prevail, they more likely to reconsider their own views, or at least to be discouraged from posting.

In sum, I expect the four-variable framework to apply much more narrowly in the Xi era rather than being totally inapplicable. Xi and the propaganda cadres under him have shown a clear preference for responsiveness through Party-controlled channels rather than microblogs, but while such channels may be effective in convincing citizens that the central government will respond to their input and take action in certain areas (such as anti-corruption), they still have limitations. Stockmann and Luo (2015) note that Sina *Weibo* (and Baidu's *Tieba*) uniquely combine both human-to-human interaction – the ability to horizontally spread information and discussion among peers or friends – with the characteristic of relying on individuals rather than news organizations as information sources. If the CCP is looking to obtain policy feedback, structured comment and report platforms that do not allow citizen-to-citizen interaction make sense. But during crises in which preventing the horizontal spread of information – whether rumor or fact – is difficult or impossible, the Xi leadership may still prefer that such interaction occur on *Weibo*, where it can be monitored, partially censored and false rumors refuted, than through the grapevine or other underground channels where greater distortion is inevitable. Thus, I expect crises in which major events grab the public's attention (and commercial media have strong incentives to report until ordered not to) to be a partial exception to the inapplicability of *responsiveness benefit* to the Xi era.

Case analysis of major online news events using *Weibo* data similar to that

in Chapters 4-6 will continue to be useful for testing the theoretical framework. However, future work also needs to test the framework's individual-level observable implications. The most important of these is that broadly speaking, co-opted channels for citizen input like the CCDI's mobile app should prove less effective *among Weibo-using demographics* than horizontal "public square" forums at increasing citizen trust in the central government or its commitment to carry out reforms. While older citizens or rural individuals who rely more on state-sanctioned input mechanisms may respond more positively to such channels, the young, urban demographics (particularly in first-tier cities) that already report the lowest levels of trust in the CCP should be skeptical of Party-sanctioned venues and place greater weight on online public pressure as the primary means of generating government responsiveness and policy change. Future survey or experimental work could test this key micro-level proposition.

7.3.3 Beyond China: 'Similar-enough' Internet-Savvy Regimes

This dissertation has focused exclusively on the Chinese case as theory-generating exercise for explaining variation in online censorship. While China is valuable to study in its own right as the world's largest authoritarian regime (and with the largest ruling party) and as a world leader in the breadth and depth of its media control techniques, sophisticated Internet interventions are increasingly a global phenomenon. Since China is unparalleled globally in terms of both the resources, and sophistication it displays in censorship, and recently has even been mentioned as a 'model' in this regard by governments interested in learning its techniques,

it can serve as a useful starting point to ask two comparative questions: a) need other countries be substantially "like China" to implement an information manipulation strategy as versatile and nuanced as China's, and b) if so, what *minimum* conditions must other countries meet (in online censorship resources, capabilities and doctrines) in order to do so? Chapter One suggested that at a minimum, other would-be selective-censoring regimes must meet three criteria: 1) large and vibrant domestic Internet companies as well as a large and active social media-using population; 2) a technologically sophisticated and functionally differentiated bureaucracy; and 3) either a single, or dominant party that does not face significant electoral competition. A corollary (but not an absolute criterion) to 3) is that such one-party states are especially likely to be able to implement a selective censorship strategy if they are organized along Leninist lines, where governing elites within the party rely on propaganda and ideology to maintain control over party members, and more broadly over society. However, potential alternatives to party ideologies do exist in practice – for example, theocracy – and could also be sufficient.

Do any countries meet all three criteria in practice? Do any even come close? The first criterion can only be met for countries with a sufficiently large domestic social media market to support the development of an indigenous Internet sector; at a minimum, the country must have tens of millions of Internet users, and be wealthy and technologically sophisticated enough to train and retain software developers in the home labor market. The country must also be illiberal enough to justify blocking or at least restricting global social media outlets like Facebook and Twitter since these have proven popular across all parts of the world and nearly all cultures in which they have gained market access. This criteron then disqualifies the overwhelming majority of countries, as either not wealthy enough (most of

sub-Saharan Africa, central Asia, and elsewhere), rapidly developing but not yet able to retain software talent in the home market (parts of Southeast Asia), or too liberal (India, Latin America, and more debatably, parts of Southeast Asia and Eastern Europe). Even for a handful of countries that are poor, technologically unsophisticated and illiberal and have the political will to block foreign websites and to censor domestically, market size may simply be too small to have a viable alternative domestic Internet sector. Cuba is a clear example of this in sharing many other similarities with China – a walled-off Internet, and a history of government information control – but would likely face great difficulty in developing a viable Internet industry even if it were wealthier.

Second, the country must have a sophisticated and functionally differentiated bureaucracy. The experience of China's Internet police suggests the hypothesis that while adequate for monitoring criminally-related online speech (broadly defined), and applying real-world coercion to bloggers, security agencies generally do a poor job of combining censorship with propaganda, or implementing sophisticated content filtering programs. Thus, I argue that country success in creating and implementing nuanced social media interventions is at least somewhat path-dependent on prior development of a media and propaganda bureaucracy. In other words, countries will struggle to develop successful 'Internet management' bureaus from whole cloth, and those with existing specialized 'state censorship' organs are much better positioned to do so. In the 20th Century, no other regime type did as much as the Leninist one-party state in developing the bureaucratic apparatus of censorship, suggesting that the Leninist model is a good reference point to which non-Leninist cases like Iran can be considered. Thus, in practice the second and third criteria are closely linked: states with functionally differentiated censorship bureaucrats tend to have a history of Leninist or quasi-Leninist organizations play-

ing a dominant role in governance and social development, as these organizations have historically proven most successful at nuanced coordination (beyond outright suppression) of the press, radio and television prior to the Internet era. Other regime types like military rule could possibly meet these two criteria (as of 2017, the Thai military junta's role in censorship comes to mind), but only to the extent that military or other organizations approximate the allocation of bureaucratic resources and responsibilities for censorship characteristic of Leninist states. While online *censorship* is definitely possible under other regime types, *selective* censorship should not be. This then casts doubt on whether states like Egypt, which have had either personalist dictatorship or military rule but did not have especially sophisticated censorship programs led by the ruling party, should be treated as comparable cases.

With these criteria in place, we are left with only a handful of potentially comparable countries with China. The only country that could completely fulfill *all* of the criteria is Russia, although at present it does not perfectly meet this standard. If it ever chose to completely block Facebook and provided it continues to experience rapid economic development, Vietnam might eventually also qualify. Finally, Iran is a potential candidate, although it is truly a borderline case as whether it meets each of the three criteria is highly debatable.[23] While detailed comparison of these countries with China is beyond the dissertation's scope, here I briefly sketch out some of the salient features of Internet control in Russia, and in Iran as a prelude to future work. I begin with Russia, which in 2016 had a population of 142 million, GDP Per Capita of $26,000, and a 73.4% Internet adoption rate, ranking seventh in the world.[24] After a move toward political liberalization in the 1990s,

[23]If pushed, one could stretch the universe of cases to several more "borderline" countries like Turkey, Venezuela, and some former Soviet states like Belarus. One example that definitely does *not* qualify is North Korea, which utterly fails the domestic market and Internet sector criterion.

[24]Source: CIA World Factbook.

Vladmir Putin's election as President in 2000 led to a gradual reassertion of central control over television and print media over the following decade, particularly following Putin's return to the presidency in 2012.

Throughout most of this period, the Russian Internet remained more open and less subject to state intervention than its Chinese counterpart. However, the situation has changed markedly in the past several years. As recently as 2009 Russia's score on Freedom House's *Freedom on the Net* report was below 50 (49 or "Partly Free" in 2009 on a scale of 0 to 100 with 100 completely "Not Free"). Since then, its score has steadily worsened, reaching 65 ("Not Free") by 2016.[25] As of 2017, Russia is clearly moving toward a much more stringent online control regime. Yet unlike China, this shift has not necessarily occurred with the cooperation or acquiescence of domestic Internet companies. Russia's largest domestic social media site is the social networking service VKontakte (VK), which was launched in 2006 and currently has 410 million active users, mainly in Russia and throughout Eastern Europe. Its founder, Pavel Durov, seems to fit the stereotype of an academically and intellectually brilliant (and fiercely independent) Internet company founder – in the mold of Mark Zuckerberg, Sergey Brin and Larry Page – and Durov indeed clashed with Russian authorities at various points during VK's early growth. In 2011, police surrounded Durov's Saint Petersburg residence after he refused to take down the pages of opposition politicians during protests over the parliamentary elections that year. Finally, in 2014 Durov was forced to resign from the Board of Directors after he refused to hand over data about Ukrainian protesters to Russian security services, and is currently living in self-imposed exile outside Russia.[26]

[25]For comparison's sake, China's score in 2016 was 88, ranking worst in the world, and Iran's was second-worst at 87.

[26]Source: The New York Times. December 2, 2014. "Once Celebrated In Russia, The Programmer Pavel Durov Chooses Exile."

While Durov's story contains biographical elements unique to him, it aptly illustrates the intensely oppositional and dissident-oriented nature of Internet space in Russia, especially when compared with the major television networks from which the vast majority of Russians get their news (Levada Center 2014). This then highlights a key difference between China and Russia. In China, Internet giants like Baidu and Tencent have "grown up" with explicit state approval, if not outright support, and founders and CEOs like Tencent's Ma Huateng are CCP members and delegates to the National People's Congress. In Russia, tech entrepreneurship has taken place more along the Western model of keeping the state at arm's length. Exploring the Russian Internet sector's relation to state authority and comparing it with China's situation is a key topic for future research, and is far beyond this section's scope. However, future research should pay attention to macro-economic and state capacity differences between the two countries/sectors. Russia lacks the sheer manpower of China's two million Internet commentators and government budgets, especially given recently depressed oil prices, may not allow for well-staffed propaganda departments or "Cyberspace Administrations" like those in China. And as a natural resource-rich state compared with China, the Russian government has much less of an incentive to promote domestic technology development. Finally, the legacy of top-down Soviet propaganda and its influence in the Putin era may simply be less well-suited to fine-grained Internet interventions (as opposed to broadcast media like television) compared with China's history of mass involvement in propaganda efforts. All that said, Putin's government *has* made moves since 2011 to shore up Internet control along Chinese lines, such as a 2014 law requiring foreign Internet companies to physically store Russian user data within Russia, or face expulsion from the market. It remains to be seen whether a declining economy and natural resource revenues, and the Internet's increasing im-

portance as a news source for citizens, will continue to incentivize the government to more aggressively intervene in online space.

Finally, the case of Iran is worth considering. Iran in 2016 had a population of 83 million, GDP Per Capita of $18,000, and a 44% Internet adoption rate, ranking 26th in the world.[27] The country's Internet market is smaller than Russia's, and Iran has no equivalent domestic network with user base and commercial success comparable to VKontakte's. In other respects, however, Iran is more comparable to China, with an equally bad score (87 to China's 88) in the *Freedom on the Net* 2016 report. And Iran resembles China in other institutional aspects, such as the creation in 2012 of the "Supreme Council of Cyberspace" (SCC), the country's top policy-making body that reports directly to Supreme Leader Ayatollah Khamenei, bypassing the executive, legislative and judicial branches (Freedom House 2016). The extent to which the ayatollahs and clerics in Iran play a role analogous to top CCP propaganda officials in China with respect to Internet censorship is a comparative question worthy of further attention. Iran is also technologically similar to China regarding content filtering techniques. Facebook and Twitter have been blocked since 2009, and website filtering is pervasive. The government also extensively uses judicial tools to enforce censorship, with the Computer Crimes Law of 2009 containing very broad language as to what constitutes an online "crime" and justifies filtering. The government is also actively promoting various domestic alternative social networks and ICT development, though so far without the same level of success as their Chinese counterparts.

Taken together, Iran's online censorship regime shares much in common with China in terms of tools, tactics and the level of repression. It may differ, however, in two respects. First, government ministries and religious authorities have

[27]Source: CIA World Factbook.

taken primary responsibility for implementing censorship rather than outsourcing this task to domestic companies. Second, Iran does not share China's emphasis on combining "negative" and "positive" (propaganda) interventions to the same extent. The state lacks any equivalent, for example, to China's "Fifty Cent Party" and has not mobilized millions of government officials in service of propaganda efforts. While the country's censorship bureaucracy is still relatively understudied and needs intensive research, based on analysis of the China case my theoretical prior is that in the absence of domestic Internet sector participation and a human effort to push "positive" interventions, Iran will be unable to implement a selective censorship program as defined in this dissertation despite Internet controls and laws otherwise comparable to China's.

Work on comparing the institutions and state-business ties that undergird different authoritarian censorship regimes has only just begun in political science. This dissertation has contributed to this nascent research program by theorizing about the resources, institutions, and ideological priors that enable states to implement unified, highly responsive and fine-tuned online information manipulation systems. Then, focusing on China, I have considered in a series of issue-specific case studies how such preconditions have enabled the CCP to implement one such system, and to adjust censorship according to well-defined shifting costs and benefits as online breaking incidents both provide top leaders with opportunities to bolster popular support, and threaten their legitimacy or even political stability. One of the most important emergent findings from this exercise has been just how many factors have been necessary for China, around the year 2012, to realize perhaps the most sophisticated apparatus in human history for shaping what citizens read and experience online during political controversies. This high bar for equating other countries' efforts to the Chinese system may ultimately mean that China

is truly peerless when it comes to Internet control, but at the very least, the work done here can serve as a basis for evaluating such a claim.

APPENDIX A

WEIBO DATA CODING PROCEDURES AND INTER-CODER RELIABILITY

A.1 Chapter Four

My co-author and I assembled a team of three coders, including the two co-authors and a third undergraduate, native Mandarin-speaking assistant. We drew a random sample of 500 posts for analysis, and worked independently to assign them into categories.[1] We then met to reconcile divergent scores according to strict rules. Through this process, we were able to agree on a consensus score for 473/500, or 94.6% of posts. As a backup procedure if consensus could not be reached but a majority was present, we broke impasses by voting in 16/500 or 3.2% of cases. Finally, in a handful of cases (11/500, or 2.2%) there were two coders who gave divergent scores, and a third who had given one or the other score (or picked a completely different category), but after discussion was "on the fence" between the other two positions; we resolved this by flipping a coin.[2] This process resulted in a set of 500 coded posts, and we report inter-coder reliability statistics in Table A.1:

[1] We later replicated the same procedure for smaller samples of 150 posts each for January 19, June 6 and June 13. Inter-coder reliability was very similar to Table A.1 below.

[2] In a handful of cases, coders gave three different scores. We followed the same procedure as above, except with simultaneous persuasion attempts in three directions. Without exception, such discussion reduced the options on the table to two codings (no cases occurred where no coder agreed to switch positions after discussion). We then followed regular rules to resolve the two-way impasse.

Table A.1: Inter-coder Reliability: Air Pollution

Statistic	Domestic vis-a-vis Foreign	Anti-Govt.	Health	AQI Monitoring
Avg. pairwise agreement	96.0%	91.2%	71.9%	94.9%
Fleiss' Kappa	0.869	0.575	0.202	0.519
Krippendorff's Alpha	0.869	0.575	0.203	0.520

Flipped coin (average across categories): 2.2%

We obtained our most reliable coding results for Domestic vis-a-vis Foreign ($\kappa = 0.869$) and middling performance for Anti-Government and AQI Monitoring. For Health, however, inter-coder reliability for these observations was low ($\kappa = 0.202$), which could certainly have greatly contributed to the null regression estimates for this measure since poor (but non-systematic) inter-coder reliability would tend to introduce attenuation bias.

A.2 Chapter Five

In contrast to similar recent projects (Cairns and Carlson 2016; Cairns and Plantan 2016) in which I developed sentiment categories and manually coded posts alongside a co-author and a research assistant (for a a three-person coding team), for this chapter I coded solo due to a lack of available resources. However, these two prior projects provided me with valuable practice in reading and coding *Weibo* posts, and increased my confidence in doing so alone this time.

I began the exercise by using structural topic modeling, or STM (Roberts *et*

Al. 2014) to identify latent scandal-related topics in the *Weibo* data and key words/phrases associated with each topic.[3] I did this by first searching the WeiboScope corpus for the simple keywords "Bo Xilai" and "Wang Lijun" and then identifying peak dates for this keyword's incidence. I took all dates with keyword counts more than two standard deviations above the year-long mean, for a total of 11 days that would later correspond to the peaks of Phases I-III. I then separated out the text from these dates and used the Txtorg program (Lucas *et Al.* 2014) to create a Term-Document Matrix, or TDM.[4] Next, I input the TDM into the above authors' topic model, implemented in the R language.[5] After model estimation, I used STM's LabelTopics function to report lists of the most frequently associated keywords with each topic.

After looking at this algorithm-generated keyword list, I then drew and read through several random samples from each of the four time periods (February 8-12, March 14-16, April 11 and September 28-29) to see which keywords were strongly associated with what I judged to be topically relevant content, and which were just noise from the automated procedure. I whittled down the list of keywords to the final ones in Section 4.1, then based on these words and holistic post reading, formally defined the sentiment categories. Next, I drew a sample of 100 'practice' posts taken evenly from across the four date ranges. After going through this coding exercise, I refined the scheme and reduced the number of categories. Finally,

[3]STM is a topic-modeling algorithm based on a family of unsupervised machine learning models called Latent Dirichlet Allocation, or LDA. See Blei, Ng and Jordan (2003) for the canonical work on this topic.

[4]A term-document matrix is a mathematical summary of the frequency of terms appearing in a text corpus, and is used as input into many natural language processing algorithms.

[5]The key researcher-chosen parameter in a structural topic model is the number of topics K. There is no "right" number of topics, but picking a nonsensically high or low number may lead to confusing or poorly interpretable results. After some experimentation, I settled on $K = 10$ topics. Later on, when I had switched from computerized topic modeling to manually reading the posts, I trimmed the number of sentiment categories ("topics") down to the five presented in this paper.

I drew a sample of 1000 posts total (250 from each date range) and proceeded to score these according to the category definitions. These posts then served both to directly estimate category proportions for those dates, and as input into *ReadMe*.

A.3 Chapter Six

My co-author, an undergraduate Mandarin-speaking research assistant, and I each independently read and scored 479 *Weibo* posts.[6] Each post contained any original text, plus any reposted or "re-tweeted" content. To simplify analysis, we counted both the original post text, and the re-posted text (if any) as part of the same message unit, i.e. we read the posts with an eye to gauge the sentiment of this overall combination, rather than considering original and re-posted sentiments separately. Below, we report common reliability statistics in Table A.2:

Table A.2: Inter-Coder Reliability: Diaoyu Dispute

Avg. pairwise agreement	60.7%
Fleiss' Kappa	52.8%
Krippendorff's Alpha	52.8%
Flipped coin	6.0%

Additionally, we calculated a unique statistic to take account of how often we had

[6]This section describes specifics of our procedure for coding the sample containing the *diaoyu-dao* keyword. With respect to the smaller sample of dispute-relevant posts not containing this keyword, the same basic procedure of reading the entire post text and assigning it to one of the eight categories was followed, except that due to resource limitations, I undertook coding alone. Although coding alone obviously precludes calculating formal inter-coder reliability, I benefited from several rounds of previous coding and team discussion. Therefore, although results are somewhat more subject to my personal biases than the team results, I am confident that they are in the neighborhood of figures that the team would have achieved.

to resort to a coin flip to break an impasse over two codings.[7] Given the difficulty of the coding exercise, about 40% of the time we resorted to brief discussion to reconcile different codings. These discussions usually lasted only a minute or two, and frequently one or more coders was eager to change his or her mind, having felt that he/she had mis-assigned a post due to error or fatigue. Instances where coders disagreed with each other to the point where arriving at a consensus score was impossible were infrequent, and occurred only 6% of the time.

After completing coding, we wished to evaluate the correspondence between our human-derived sentiment categories, and keyword proxies. One measure of this is what percentage of posts containing a given keyword ended up belonging to the "appropriate" category for that keyword.[8] This measure is in Table A.3 below:

[7] In situations where the three coders each assigned three separate scores to a post, and remained at deadlock after discussion per the above rules, we flipped a coin twice. This situation was rare and only occurred a few times.

[8] Benchmarking our keyword measures' ability to proxy for underlying categories in this manner is analogous to the "precision" measure in the computer science literature – we are more concerned about false positives than about keywords' ability to retrieve all relevant content for a category. We are aware that using keywords we cannot infer fluctuations in sentiment categories from changes in keyword counts over time. However, as a qualitative as well as quantitative illustration of the sorts of sentiments prevalent during the dispute, we believe our keyword approach to be a valuable complement to the directly estimated category proportions, as well as the ReadMe results.

Table A.3: Percent of Posts with a Given Keyword Belonging to "Correct" Category: Diaoyu Dispute

Term	Correct category	(Posts with category plus keyword)/(posts with keyword)	% of posts
Anti-government (*tianchao*)	5	9/9	100
"Boycott Japanese goods" (*dizhi rihuo*)	4	9/19	47
"The Diaoyu islands are China's" (*diaoyudao shi zhongguode*)	2	23/48	48
"Smash" (Japanese cars, etc.) (*za* (*che*))	1	18/26	69
"Rational" (patriotism) (*lixing*)	1	6/8	75
Total		65/110	59

The above results show that the incidence of *tianchao* perfectly predicts a post belonging to the anti-government category, and the keywords *za* and *lixing* proxy moderately well for the moderate category.

APPENDIX B

CORRECTING FOR DATA COLLECTION BIAS IN ESTIMATING CENSORSHIP

One of the major difficulties in using Chinese social media data is how to deal with the bias induced by state censorship, since researchers attempting to "harvest" such data are only able to observe the blog posts they are able to download faster than censors can delete these posts. However, as long as researchers are able to capture a fraction of all censored posts, it may be possible to estimate the true censorship rate. First, assume that out of the sample of around 43,000 Weibo bloggers, some fraction decide to write a post in response to some event.[1] Also assume that individuals who choose to write a post do so immediately following the event.[2] What I want to know is how many of these posts will survive (not be censored) long enough to appear in the WeiboScope data. This information is necessary to calculate my primary quantity of interest – the true censorship rate:

$$R_{true} = \frac{C_{obs} + C_{hid}}{C_{obs} + C_{hid} + P} \tag{B.1}$$

Where R_{true} is the true rate, expressed as the proportion of censored over total posts, P is posts that are never censored (all of which appear in the dataset), C_{obs} is the number of posts marked as "censored" in the dataset, and C_{hid} is those posts that get censored, but do not appear in the dataset because they are deleted

[1]The fraction that decides to write versus not write a post in response to breaking political news does not matter for modeling the data-generating process, and I do not consider it further because I only care about generalizing my findings to those individuals who do post – I do not seek to explain "participation" in the dataset.

[2]While this simplifies reality, the findings of the exercise here generalize easily to cases where individuals choose different durations after an event at which to write a first post, provided that posts occurring later on follow the same censorship distribution over time as their immediate counterparts.

sooner than the Hong Kong team can download the *Weibo* user timelines that contain them. The WeiboScope data scraping process, as described in Fu, Chan and Chau (2013), involved periodically returning to the pages of the 43,000 users, downloading a copy of the timeline each time. If a post got deleted between crawls (i.e. after the team's program had crawled a page during a particular iteration, but before the next one), then the researchers could compare the new record to the old one, identify the post that had disappeared in the interim, and mark it as censored. However, due to limits set by Sina.com, the team could only crawl most of these pages (38,000 out of the 43,000, who constituted the "Verified" user group) once every 24 hours. Given a uniform distribution of sensitive post-inducing events (i.e., that they were equally likely to occur over a given 24-hour period), the average time between when a post would go up and when that Verified user's page would be crawled, would be 12 hours. Since Zhu et al. (2013) find that most censorship occurs within an hour or so of the post time, most censored posts from the Verified users were unlikely to make it into the dataset. Thus, the dataset is truncated, and R_{true} will be biased. What I have is the observed rate, R_{obs}, in Equation B.2:

$$R_{obs} = \frac{C_{obs}}{C_{obs} + P} \tag{B.2}$$

Since C_{hid} is missing, $R_{obs} < R_{true}$, i.e. the observed censorship rate is biased downward. But how much so? The observed year-long average rate for the topics in Chapters 4-6 is between 12% and 17%, an oddly low figure given that other studies (King, Pan and Roberts 2013) have measured the true rate during sensitive events to be closer to 60% and I have no reason to think that Chapters 4-6 are any exception. To calculate the true rate, I need to know the true number of

posts censored, N_{true}, which is related to N_{obs}, the number of censored posts that I actually observe, via some probability distribution that models the speed with which censors remove posts during sensitive episodes.[3]

Since I do not know the true distribution, I need to look for an empirical example that provides a good approximation. The best available so far is the finding by Zhu *et Al.* (2013), who note that "nearly 90% of deletion events happen within the first 24 hours" (p. 1). Conveniently, this time window is the same as that of the unbiased portion of the data: 100 percent of posts will be observed, and correctly identified as censored or not, if they survive 24 hours or more. Since *et Al.* found that 90 percent of censorship occurs before 24 hours, 10 percent must occur after, sometimes days or weeks later. Since I observe this 10 percent, and critically, assuming that the form of the censorship distribution over time is the same in my data as in that of *Zhu et Al.*, the ratio of what I observe to what gets missed must be 1:9, e.g. $C_{hid} = 9C_{obs}$. This suggests that multiplying C_{obs} by a factor of 10 will get me close to the true rate. Plugging this into Equation B.2 gives:

$$R^{*}_{true} = \frac{C_{obs} + 9C_{obs}}{C_{obs} + 9C_{obs} + P} \tag{B.3}$$

B.1 Chapter Four

Applying this equation to Chapter Four gives Table B.1 below, which shows the observed censorship rate, the number of observed posts (including non-censored

[3] An earlier version of this appendix for Cairns and Carlson (2016) contained a mathematical formalization, omitted for space purposes.

posts), and the estimated true rate. I calculate these numbers for the posts based on my pollution-relevant keywords, and then estimate the joint rate among all topic-relevant posts:

Table B.1: Observed Versus True Censorship Rates For Peak Discussion Dates: Air Pollution (90%/24 hrs)

Date	Posts	Observed rate	True rate
1/2–6/5	181 (avg)	.11	.49
6/6	1460	.20	.71
6/13	2363	.04	.30
6/14–12/30	164 (avg)	.18	.64
Year	71,088	.15	.57

Given that I am applying another *Weibo* study's findings to a different dataset, the question might arise, given that my data consist of journalists, dissidents, and Verified users with more than 10,000 followers – all sensitive groups in censors' eyes – whether 90 percent within 24 hours is too slow a rate for the sample. *Zhu et al.* and King, Pan and Roberts both find that some small fraction of ultimately censored posts typically linger for days after an incident – the question here is how much. My main empirical concern in this paper is under-, not over-estimating the censorship rate. If I assume that the true number is 95 percent within 24 hours, i.e. $C_{hid} = 19C_{obs}$ then plugging these numbers into Equation 2 yields Table B.2:[4]

[4]Given that the assumption of 90% post deletion within 24 hours is already very pessimistic, I believe 95% represents an absolute worst-case scenario. 90% within 24 hours would only be true if the entire year of 2012 were constantly filled with sensitive pollution-related online outbursts – it is unlikely that the in-house censors Sina employs to delete posts devote the resources and attention necessary to achieve such a fast deletion rate for non-critical events, although Appendix B.1 does explore this further. This is why I think my adjusted measure of censorship probably overestimates the true rate for much of the year. However, since for statistical purposes I am

Table B.2: Observed Versus True Censorship Rates For Peak Discussion Dates: Air Pollution (95%/24 hrs)

Date	Posts	Observed rate	True rate
1/2–6/5	181 (avg)	.11	.61
6/6	1460	.20	.71
6/13	2363	.04	.46
6/14–12/30	164 (avg)	.18	.76
Year	71,088	.15	.70

The rates above are higher than the previous estimates. However, the estimated mean censorship rate (for keyword posts) for June 13 still falls below 50%, a rate far less than surrounding dates.

B.2 Chapter Five

Applying Equation 3 to Chapter Five gives Table B.3 below:

primarily concerned with censorship fluctuations rather than the level, the specific censorship adjustment I choose should have little impact on my results.

Table B.3: Observed Versus True Censorship Rates For Peak Discussion Dates: Bo Xilai Scandal (90%/24 hrs)

Date	Posts	Observed rate	True rate
2/8	1745	.04	.28
2/9	3528	.03	.22
2/10	3754	.03	.23
2/11	5765	.01	.11
2/12	1462	.04	.30
3/14	1203	.11	.56
3/15	4338	.10	.52
3/16	1441	.09	.49
4/11	2385	.06	.40
9/28	1235	.18	.69
9/29	1114	.19	.70
All Phases (avg)	394	.17	.61

Next, Table B.4 below assumes the true number is 95% within 24 hours:

Table B.4: Observed Versus True Censorship Rates For Peak Discussion Dates: Bo Xilai Scandal (95%/24 hrs)

Date	Posts	Observed rate	True rate
2/8	1745	.04	.43
2/9	3528	.03	.37
2/10	3754	.03	.37
2/11	5765	.01	.20
2/12	1462	.04	.47
3/14	1203	.11	.71
3/15	4338	.10	.68
3/16	1441	.09	.66
4/11	2385	.06	.57
9/28	1235	.18	.82
9/29	1114	.19	.82
All Phases (avg)	394	.17	.73

The rates above are higher than the previous estimates. However, the estimated mean censorship rate for February 8-12 (Phase I) still hovers around 40% and goes as low as 20%, a rate far less than later dates.

B.3 Chapter Six

Applying Equation 3 to Chapter Six gives Table B.5 below. I calculate these numbers for the *diaoyudao* keyword sample, and the non-keyword sample (for key dates), and then estimate the joint rate among all topic-relevant posts:

Table B.5: Observed Versus True Censorship Rates For Peak Discussion Dates: Diaoyu Dispute (90%/24 hrs)

Date	Posts (with keyword)	Estimated posts (w/o keyword)	Obs. rate (*diaoyu-dao* keyword)	True rate (*diaoyu-dao* keyword)	True rate for posts w/o keyword (95% CI)	Rate for whole population (95% CI)
8/15	6173		9.8	52.0		
8/16	6427		8.8	49.0		
8/17	3581		8.4	47.9		
8/18	7044	23,137	3.2	24.7	0-38.5	6.1-36.1
8/19	7180		7.2	43.7		
8/20	2411		11.4	56.2		
9/9	680		11.2	55.7		
9/10	3565		14.0	62.0		
9/11	10130		14.2	62.3		
9/12	8480		13.4	60.8		
9/13	7053		12.7	59.3		
9/14	9455		13.0	59.8		
9/15	7255	24,701	13.3	60.5	51.0-82.5	53.0-79.5
9/16	6243		12.6	58.9		
9/17	5873		13.4	60.7		
9/18	7394		12.6	59.0		

Next, Table B.6 assumes 95% within 24 hours:[5]

[5]A full defense of this assumption is beyond the chapter's scope, but here I briefly describe my logic. I think of 95% as a very conservative upper bound according to the following: if the true amount within 24 hours were indeed 95%, this would imply that a very large volume of Diaoyu-relevant *Weibo* content was created by users, and then wiped out of existence before being captured in the dataset. Comparing this potential volume with post surges from *Weibo's* top topic in 2012 (the London Olympics), I set a 'face validity' limit to how large the pool of deleted posts could have been, and therefore an upper limit to the maximum percent deleted within 24 hours. For example, I count 47,821 as the number of posts in WeiboScope containing the keyword "Olympic" (*aoyun*, or *aolinpike*) on July 28, 2012, the date that the Opening Ceremony for the London Olympics was broadcast Beijing time. Then I assume, using the above figure, that this keyword was heavily censored as if it proxied for a collective action topic (a dubious, worst-case assumption given that most discussion about the Olympics was surely non-political), and I use *Zhu et Al.'s* 90% estimate in extrapolating and 'adding back' a large hypothetical number of censored posts. The phrase "2012 London Olympics" (2012 *nian lundun ao yun hui*) was *Weibo's* top trending topic of 2012 according to Sina.com; in comparison, "The

Table B.6: Observed Versus True Censorship Rates For Peak Discussion Dates: Diaoyu Dispute (95%/24 hrs)

Date	Posts (with keyword)	Estimated posts (w/o keyword)	Obs. rate (*diaoyu-dao* keyword)	True rate (*diaoyu-dao* keyword)	True rate for posts w/o keyword (95% CI)	Rate for whole population (95% CI)
8/15	6173		9.8	68.4		
8/16	6427		8.8	65.8		
8/17	3581		8.4	64.8		
8/18	7044	23,137	3.2	39.6	0-55.6	11.6-53.1
8/19	7180		7.2	60.8		
9/9	680		11.2	71.6		
9/10	3565		14.0	76.5		
9/11	10130		14.2	76.8		
9/12	8480		13.4	75.6		
9/13	7053		12.7	74.5		
9/14	9455		13.0	74.9		
9/15	7255	24,701	13.3	75.4	67.6-90.4	69.3-88.6
9/16	6243		12.6	74.2		
9/17	5873		13.4	75.6		
9/18	7394		12.6	74.2		

The rates above are higher than the previous estimates. However, the estimated mean censorship rate (with the *diaoyudao* keyword) for August 18 still falls short of 40 percent, a rate far less than in September and lower than that for other collective events.

Diaoyu Islands Are China's" (*diaoyudao shi zhongguode*) ranked tenth. If I allow that the total number of pre-censorship Diaoyu-relevant posts on August 18 could not have been greater than the Olympics-related figure above, e.g. $Diaoyuposts < 47,821$, then for this inequality to hold, the percent within 24 hours could not have exceeded about 91.45%. Given this, I think that an estimate of 95% is exceedingly high, going well beyond a more feasible maximum; I choose this high number to demonstrate the robustness of my results subject to all assumptions presented here.

B.4 Face Validity Checks for the Censorship Measure

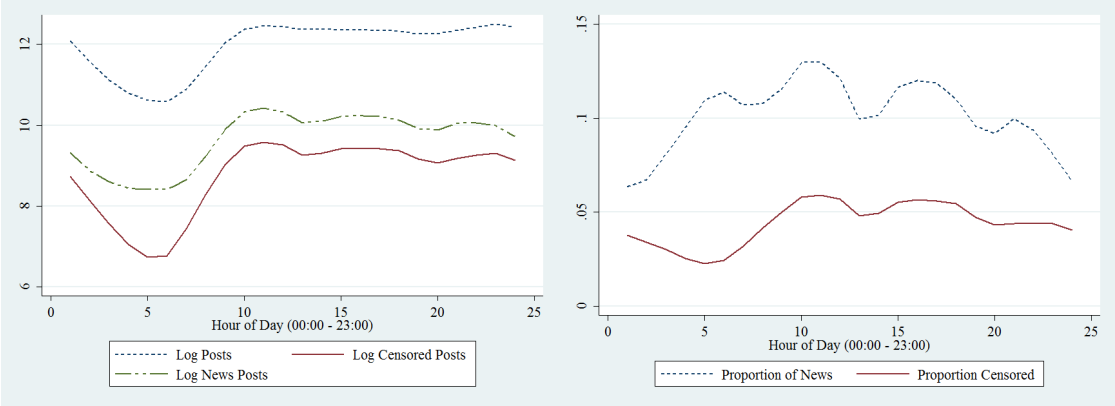
Throughout the dissertation, I have relied on a relatively simple measure of censorship: the fraction of posts marked "deleted last seen" over total posts per day – adjusted by the formulas in Appendix B. While in Chapters 4-6 this measure does indeed appear to respond to real-world events in ways that theory might expect – for example, sharply increasing or falling in response to breaking news – room for doubt may remain as to the representativeness of these cases for measure validity. Beyond the incidents in Chapters 4-6, does the measure more generally behave in ways that might be expected? For example, is censorship in the *WeiboScope data* less during nights and weekends, as news interviews with Sina employees have indicated?[6] And is censorship low during *non*-sensitive "hot topics" like discussion of television premieres? In this appendix, I show evidence that the censorship measure responds distinctly and immediately to predictable real-world events, like street protests and natural disasters with the potential to foment anti-government discontent. Just as importantly, censorship of non-political topics varies less and is much lower than for sensitive events.

The following examples are meant to be clear-cut rather than borderline examples of (non-) censorship. I begin first, however, with graphing the measure's basic properties over 24-hour and weekly cycles, and during a major holiday (the Chinese New Year). Because censorship is closely related to news reports on *Weibo* and because there is less news at night and on weekends and holidays, we should expect less censorship during all three times. Figure B.1 shows the 24-hour average for all *Weibo* activity in the sample, calculated over the whole year:[7]

[6]Source: Reuters. September 11, 2013. "At Sina Weibo's censorship hub, China's Little Brothers cleanse online chatter."

[7]All censored post counts and censorship rates in Appendix B.1 are the *pre-adjustment* rates.

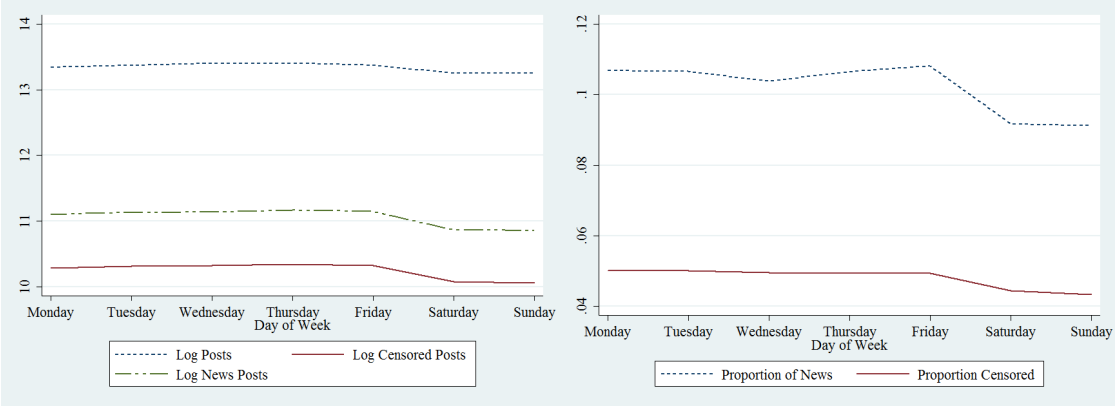
Figure B.1: *Weibo* Post Volume/Proportions: 24-hour Activity (Year-long Average)



As might be expected, the left panel shows a dip from midnight to 5am in log posts, log news posts (calculated as the count of posts containing a left bracket ("[") which signifies a news link), and the log censored post count. In the right panel, even as more news story links begin to appear on *Weibo* in the morning's early hours, the censorship rate continues to decline until 5am. Logically, this makes sense: the fraction of news content on *Weibo* goes up at night because some journalists remain hard at work publishing and then social media-posting stories, while most *Weibo* users are still asleep and so not re-tweeting the links. Note that for the rest of the day after 5am, total posts, the fraction of news posts and the censorship rate all follow the same pattern: increasing until just before noon, an early afternoon dip, a late afternoon dip and then a gradual decline. Next, Figure B.2 shows the average weekly pattern:

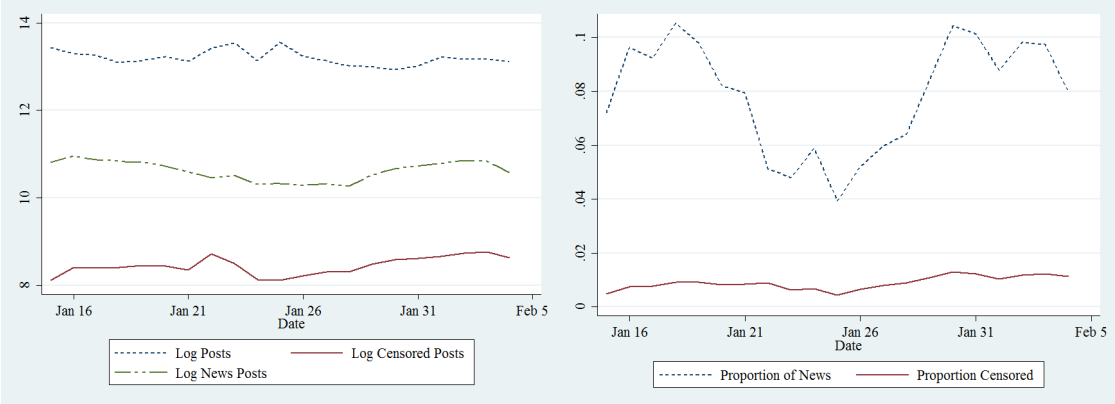
This is the only part of the dissertation that does *not* apply Appendix B's correction. I show the pre-adjustment rates in order to more faithfully represent the *WeiboScope* data's actual variation. Exactly how much downward bias the raw rates have does not really matter since I care about the changes/fluctuations more than the exact levels.

Figure B.2: *Weibo* Post Volume/Proportions: Weekly Activity (Year-long Average)



Post, news post and censored post levels are steady Monday-Friday, but then noticeably decline over the weekend. The proportions of news and censored posts similarly decline on Saturday and bottom out on Sunday. A more interesting pattern obtained in 2012 over the Chinese New Year, shown in Figure B.3:

Figure B.3: *Weibo* Post Volume/Proportions: Before, During and After Chinese New Year (Jan 15-Feb 4; New Year on Jan 23)



I sampled the days January 15 - February 4 to cover the entire one-week New Year holiday, as well as several days before and after (Chinese New Year, or *chunjie*

itself fell on January 23 that year according to the Lunar calendar, and most Chinese get the entire following week off). New Year's Eve (January 22) saw a brief uptick in posts and censored posts due to general discussion about the new year and about the annual controversy about whether private fireworks in major cities – which are traditional but substantially increase air pollution and have led to injuries and deaths – should be allowed or banned by city governments.[8]. Otherwise, the right panel shows that the censorship rate, which was already quite low before the New Year, declined even further around January 22/23 and stayed very low until about January 24, when it gradually increased. Clearly, however, it did not spike at any point despite considerable variation in the amount of news posts.

Having observed that daily/weekly and holiday patterns of posts and censorship fit basic expectations about their behavior, I now turn to five specific incidents that occurred during 2012, each of which has an expected censorship (non-) response. The first two are also analyzed in Zhu *et Al's* 2013 paper, and serve to compare the *WeiboScope* measure to their findings.[9] Figure B.4 shows posts, news and censorship related to the public reaction to a torrential rainstorm in Beijing over the weekend of July 21-22 that left nearly the entire city flooded, and caused major property damage and dozens of deaths. I sampled all posts from the dataset that contained both the keyword "Beijing" and "rainstorm" (Mandarin: *baoyu*). City officials failed to warn the public via news and online channels until late Saturday, after the rainstorm had already been underway for several hours. Additionally, despite rapid urbanization and development since the 1990s, the city's drainage

[8]I know this due to the post sampling and reading my co-author and I did for January 22 as part of Chapter Four's air pollution study

[9]Zhu *et Al* find that the keywords "Beijing rainstorm" and "Qidong" survived among the shortest time compared with the basket of keywords they analyzed. A post containing "Qidong" lasted just 1.18 hours, while "Beijing rainstorms" lasted 2.65 hours on average (see authors' Table 3).

system was not up to the task. Occurring as it did in China's capital and peak *Weibo*-using city, the disaster typified the sort of potentially politically sensitive incident likely to draw censorship on *Weibo* as angry citizens complained about local officials' and the city's lack of preparedness.

Figure B.4: *Weibo* Post Volume/Proportions: Beijing Rainstorm (Jul 21-29)

The left panel shows two peaks in post activity: one immediately after the disaster on July 22-23, and another some days later as its aftermath and the city government's poor response became clear.[10] The right panel shows an initial *decline* in the censorship rate, possibly because censors at Sina were waiting for government guidance about how to manage related online commentary. This is also consistent, however, with *visible censorship cost* being high: occurring as it did in Beijing and affecting all residents including many central government officials, the disaster was impossible to hide, which meant that too-aggressive censorship could have backfired if imposed quickly. As news reports raised questions about the government's role on July 24 and 25, the censors kicked into action; however,

[10]Source: South China Morning Post. July 26, 2012. "'A killer storm but no one warned us'; Four days after record deluge, full extent of devastation has yet to be revealed and death toll has not been updated since Sunday, but officials deny cover-up"

censorship only maxed out at about 0.1.[11] The key point, though, is that censorship closely followed the trend in increased news reporting on July 24-25. Next, Figure B.5 considers an incident involving street protests in Qidong, Jiangsu province over the proposed construction of an industrial waste pipeline (posts containing keyword "Qidong"):

Figure B.5: *Weibo* Post Volume/Proportions: Qidong Protests (Jul 27-Aug 2)

On July 28, about 1,000 protesters took to the streets, storming the city government office and forcing the mayor to wear a protest t-shirt.[12] This action led to the project's indefinite cancellation, but despite its success was heavily censored on *Weibo*. Such censorship was likely for a few reasons: the protest occurred in Qidong, which is only an hour's drive north of major *Weibo*-using city Shanghai, and thus was likely to attract attention. It involved real-world (and even rough and violent) collective action, and had the potential to generate national resonance as an example of a NIMBY protest. Indeed, the left panel shows a rather large surge

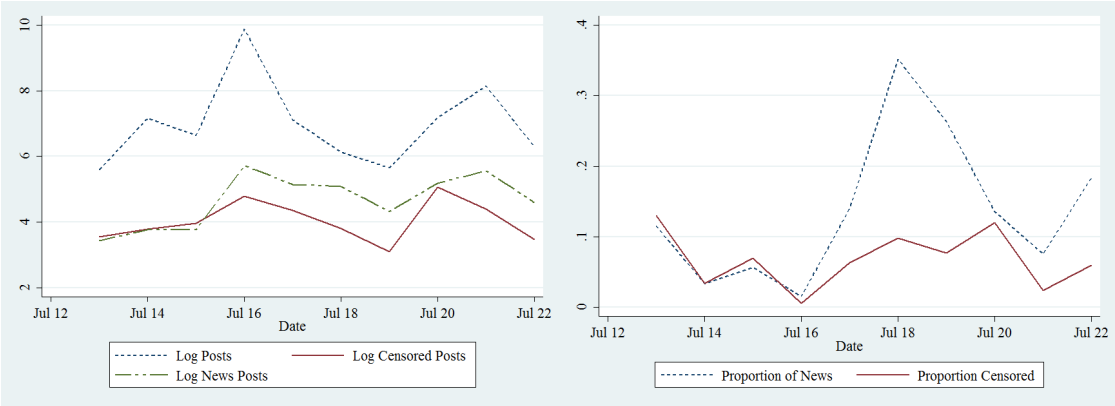
[11] With a volume of about 3,000 observed posts on July 24, this would translate, according to Appendix B, into an *adjusted* rate of around 53%, which is still a bit lower than the adjusted peak censorship rates in Chapters 4-6.

[12] Source: South China Morning Post. July 29, 2012. "City scraps waste pipeline after thousands protest; Party boss has shirt torn off his back as crowd of demonstrators invades local headquarters."

in posts on the day of the protest ($N = 2,425$, or just under e^8) with sustained activity until around July 31. Again, perhaps due to Sina officials' waiting on orders, censorship is only moderately high on July 28 (about 0.1) before increasing sharply thereafter. Still, the rise in censorship again coincides with increased news coverage in the days after the protest. And censorship declined on July 31 only because overall post volume dropped sharply on that date, which could be read as a sign of censor success in putting out the fire. Both this example and the previous one show that although the censorship rate's "natural" real-world behavior may sometimes involve a 1-2 day lag for government officials to react, after that censorship is swift and sustained.

In contrast to the above, the next two figures depict cases where spikes in censorship would not normally be anticipated. Figure B.6 shows posts and censorship in response to the premiere of popular singing competition "The Voice of China" (posts containing the show's name: *zhongguo hao shengyin* in Mandarin), which was one of *Weibo's* most-discussed topics of 2012 and premiered Season One on July 13.

Figure B.6: *Weibo* Post Volume/Proportions: "Voice of China" Premiere (Jul 13-22)

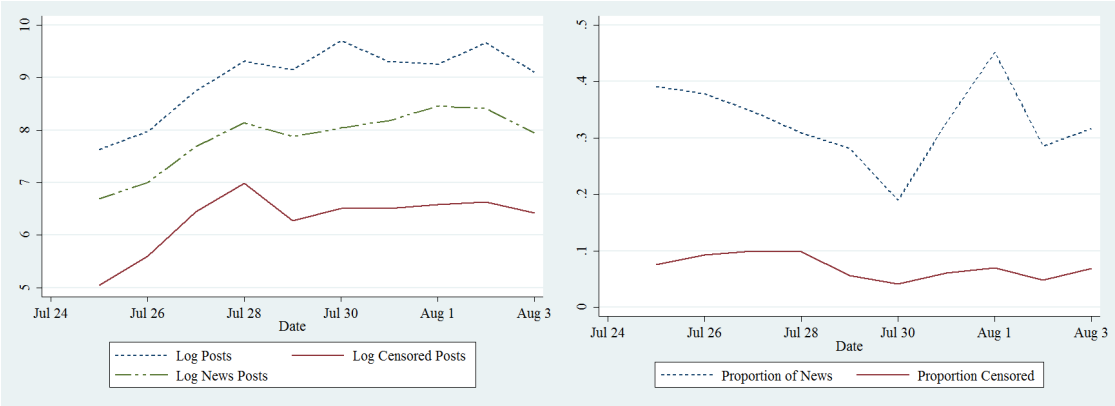


Looking at the left panel, it is somewhat surprising that the count of censored posts rises at all, given that Voice of China was widely popular and discussed, mostly without controversy, in a variety of media. However, the right panel shows that the censorship *rate* clearly drops after the show's premiere. Later, on July 16 the proportion of News posts surged, accompanied by an increase in censorship. Detailed inquiry into the case and a reading of post samples would be needed to determine what exact news (and netizen commentary) drew censors' attention. That said, censorship still remained relatively low (below 0.1) in the 10 days following the premiere. One potential controversy may have revolved around the fact that the show involved a form of democratic selection where expert panelists and ultimately media professionals were allowed to vote for their favorite singer during various rounds – possibly a bit sensitive in a one-party Communist state. However, in the absence of more detailed analysis, the graph patterns are still generally consistent with the theoretical expectation of lower and less "bursty" censorship in an ostensibly non-sensitive topic.

A second example of a "non-political" event was the 2012 London Olympics, which was *Weibo's* Number One trending topic of 2012 (I sampled posts containing either *aoyunhui* or *aolinpike*, both meaning "Olympics" in Mandarin).[13] The Opening Ceremony was broadcast Saturday morning Beijing time on July 28, an inflection point evident in Figure B.7:

[13]Source: China Internet Watch. https://www.chinainternetwatch.com/1899/top-15-most-popular-topics-on-weibo-2012/. Accessed March 11, 2017.

Figure B.7: *Weibo* Post Volume/Proportions: London Olympics (Jul 25-Aug 3)

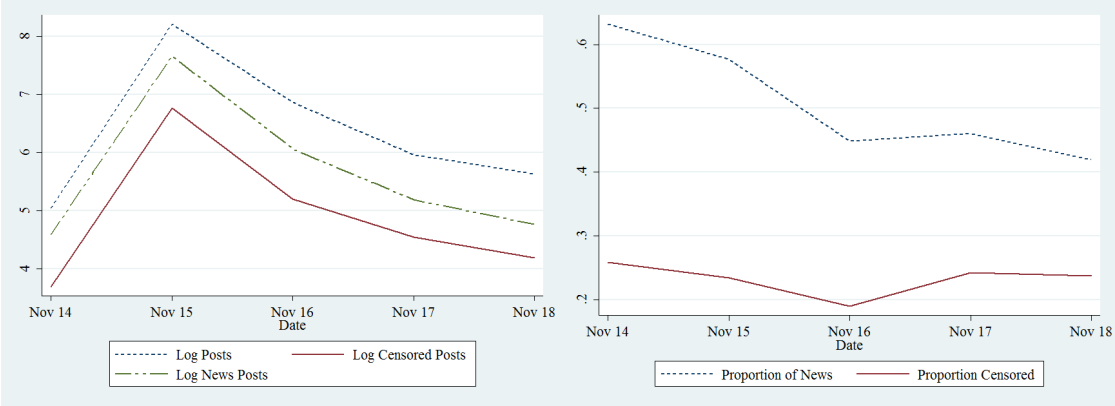


Similar to other graphs, the left panel shows a steady increase in posts, news posts and censored posts until the 28th, when censorship declines even as the daily post count increases further. More strikingly, in the right panel the proportion of news posts surges after July 30 – around the time when the events (like diving) in which Chinese athletes often do well were featured on national TV networks.[14] Yet the censorship rate remains well below 0.1 over the next several days, confirming the expectation that the vast majority of *Weibo* chatter about the Games was not politically sensitive.

Finally, I highlight a unique, politically sensitive event where theory would reasonably predict censorship to be high and then suddenly fall: the 18th Party Congress and the official selection of Xi Jinping (keyword: *Xi Jinping*) as Party Secretary and head of state in November. Figure B.8 shows the behavior of posts that mentioned Xi:

[14]Of note, the diving events in which the Chinese team is dominant began on July 29, and China won its first gold of the Olympics that day in the Women's Synchronized 3m Springboard.

Figure B.8: *Weibo* Post Volume/Proportions: Keyword "Xi Jinping" during and after 18th Party Congress (Nov 14-18)



The left panel clearly shows mentions of China's new top leader peaking on November 15, the day that Xi's official ascension and the rest of the Politburo Standing Committee lineup were publicly announced. However, the censorship rate (along with the news rate) starts falling on November 14, one day before the announcement despite being generally high during the preceding week; in the *WeiboScope* data, on November 11-13 there were a *total* of just 11 posts that mentioned Xi. This extremely low number does not indicate a lack of interest but much more likely very aggressive and rapid keyword filtering and deletion by censors (before *WeiboScope* could record them) of any posts containing the future leader's name, given widespread knowledge (and previous media coverage) that Xi would be the one chosen, and the occurrence of what many considered China's most important leadership transition in decades. In contrast, November 14 had 155 posts, increasing to 3,697 on the 15th.[15] There appear to have been deliberate instructions given to censors to block Xi's name prior to the new Standing Committee announcement,

[15]In cases where an extremely low post count in the single or low double digits (or zero count) is observed in *WeiboScope* for an event that should have prompted massive attention, it is not possible to estimate the "true" censorship rate according to Appendix B because there are not enough posts to extrapolate the quantity of 'missing' censored posts from the observed ones.

then unblock it (and the names of other PBSC officials) once they had ascended to power; indeed, a study of keyword blocking during the 18th Congress by Ng and Landry (2012) directly and strongly supports this claim.[16]

As of 2012, the censorship measure contained in Fu, Chan and Chau's *Weibo-Scope* data represented a cutting-edge attempt to measure Sina *Weibo* censorship in an extremely challenging data-gathering environment, in which both Sina and the Chinese government later took steps to make large-scale downloading of post data even more difficult. While the measure has significant limitations, the findings in this dissertation have not relied on its precision in estimating the exact censorship rate (or in capturing all forms of censorship: see Chapter Seven), but rather in its ability to rise and fall consistently according to real-world censorship-triggering events. The above figures as well as those in Chapters 4-6 support this premise by showing the measure's responsiveness to a range of circumstances and examples in which other sources – news reports, and other *Weibo* samples from 2012 – provide independent evidence of the censors' behavior, and by drawing on other studies (King, Pan and Roberts 2013; 2014) that strongly suggest instances in which censorship should spike. While continued innovation is called for in the "cat and mouse" game of beating the censors to the data, these examples should suffice to increase confidence that the dissertation's response variable has been adequately measured.

[16]Ng and Landry run "daily searches on the names of all 2,270 delegates to the Party Congress on Sina Weibo for five weeks before and after the event" (p. 1), and find (see Figure 4 of their paper) that Xi's name as well as those of all 10 re-elected Politburo officials were blocked prior to November 15, with multiple officials' names unblocked around the time of the announcement. It is encouraging to see this result in separately collected *Weibo* data from the same time period.

APPENDIX C

CHECKING ESTIMATE RELIABILITY FROM THE *READ ME* COMPUTER-ASSISTED TEXT ANALYSIS (CATA) METHOD

ReadMe is a computer-assisted text analysis (CATA) method developed by Daniel Hopkins and Gary King (2010) and tailored for a content analysis task often of use to social scientists: how to estimate the proportion of documents, social media posts, or other forms of text data belonging to a certain category within an overall corpus. The authors discuss two alternatives to their procedure for this task: hand-coding many documents, and the computer science technique of machine learning. Regarding hand coding, say researchers' goal is to estimate, for example, the proportion of blog posts critical of the Chinese government taken from a large sample of social media text. If investigators want to estimate the proportion for several different sub-samples (such as dates, or groups of bloggers) within this corpus, they must draw and hand-code one statistically robust sample from each sub-group, a feat potentially involving the laborious reading and scoring of thousands or more posts and exceeding the time and human resources of many scholars. This has led researchers to look into automated methods such as machine learning; however, Hopkins and King argue that the machine learning approaches currently being developed in computer science are poorly suited to the task of estimating category proportions, being tailored instead toward the classification of individual documents.

Even if machine learning algorithms are able to classify individual posts with high accuracy, they can be biased estimators of the aggregate proportions. In response, Hopkins and King's method produces approximately unbiased proportions for sample sizes of as little as about 500 posts. More importantly, Hopkins

and King claim that as long as the desired sub-samples are linguistically similar enough, one can "train" the algorithm on a single sub-sample and then use it to estimate out-of-sample category proportions for the other sub-groups. The method's success rests on only three factors. First, human coders are necessary to score an initial "example" sample of posts ("training data" in computer science jargon), and they should be able to define exhaustive, mutually exclusive categories with a high degree of inter-coder reliability. Second, all sub-samples within the corpus to which the categories are to be applied should be linguistically similar – that is, they should use the same *set* of terms to describe, say, negative sentiment toward the Chinese government; however, the exact frequency of different words people use is allowed to vary across sub-samples. Third, *ReadMe* works by randomly drawing sub-sets of "features" (words) from the human-sorted example documents and using this information to estimate the proportions, and researchers must define a parameter N of the number of example words to be viewed with each draw.[1] For *ReadMe* to reliably estimate the out-of-sample proportions from the training data, researchers' chosen N value must be within a few integers of the optimal value.[2]

C.1 Chapter Four

The challenge I encountered for this paper was how to get reliable estimates

[1]There also exists a second parameter called the "threshold" that tells *ReadMe* to ignore in its calculations posts that occur in less than X proportion or greater than $1 - X$ of documents. The idea is to avoid *ReadMe* attempting to calibrate itself on "noise", such as the mention of an individual's name that occurs repeatedly throughout the episode but has little predictive value for the sentiment categories (such as "Bo Xilai" in Chapter Five). In practice, in Chapters Four and Five I found that changing this threshold from the default of $X = .01$ only weakened *ReadMe's* performance, so I chose the default. In Chapter Six I set $X = .005$.

[2]While values for N have no theoretical upper bound, in practice they are almost always between about 4 and 25 words.

of the sentiment measure proportions for 364 days (to generate year-long time series) without having to code more than a thousand or so posts. In developing the training data set for input into *ReadMe*, I strove to balance posts that would represent the year overall with ensuring that the training set contained enough relevant examples from my key dates during the year in which discussion surged. In the end, my co-author, our undergraduate research assistant, and I ended up coding a random sample of 500 posts taken from the entire year, and augmenting the training data with additional samples of 150 posts from June 6 and 13, and one additional date (January 19) that also saw a large topic-relevant surge.[3] Hand-picking these dates also gave us directly estimated proportions for them, which provided a basis for comparison with the algorithm results; in other words, we were able to check, and fine-tune *ReadMe's* performance by measuring its classification error for those dates.[4] After comparing values for N from 4 to 24, we found that 16 features yielded the lowest root mean squared classification error across the three hand-coded dates.[5] Table C.1 shows results for the four hand-coded measures:

Errors vary widely across dates and categories but are generally greater for June 6 and 13 than for January 19; I attributed this discrepancy to the exceptional nature of the two dates, which contained speech patterns that stood out from the rest of the year. Through a close reading of posts, I found January 19 to contain

[3]This gave us a total of 950 posts in the training set. To simplify *ReadMe's* task and because it yielded better results, we treated the identification of each sentiment measure proportion as a binary task (a post either did, or did not belong to that measure) rather than attempting to have the algorithm do a multinomial classification of all measures at once (Chapters Five and Six do involve a multinomial classification).

[4]Since Chinese text contains no white spaces between words, using a character segmentation algorithm is necessary to separate out individual words before inputting into *ReadMe*. For chapters Four and Six I used the MMSEG algorithm (http://technology.chtsai.org/mmseg/), while for Chapter Five I used the more modern Stanford Word Segmenter (http://nlp.stanford.edu/software/segmenter.shtml). Both algorithms perform very similarly, and the choice between them is unlikely to affect results.

[5]$N = 16$ also happens to be *ReadMe's* default value, suggesting that this value is a good fit for many applications.

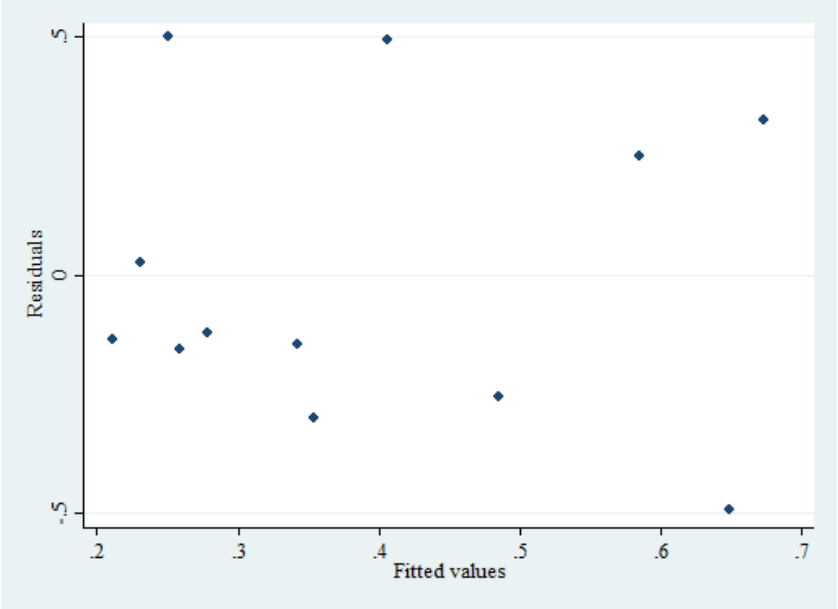
Table C.1: *ReadMe* Versus Hand-Coded Estimates: Air Pollution

Date	Measure	Dom.v.For.	Anti-Govt	Health	AQI	Monitoring
1/19	*ReadMe*	0.104	0.056	0.198	0.157	
1/19	Hand-coded	0.1	0.26	0.24	0.133	
1/19	**Error**	**0.004**	**0.204**	**0.042**	**0.024**	
6/6	*ReadMe*	0.159	0.23	0.75	0.257	
6/6	Hand-coded	0.753	0.48	0.087	0.053	
6/6	**Error**	**0.594**	**0.25**	**0.663**	**0.204**	
6/13	*ReadMe*	1.0	0.837	0.9	0.077	
6/13	Hand-coded	0.793	0.647	0.347	0.02	
6/13	**Error**	**0.207**	**0.19**	**0.553**	**0.057**	

$Num.features = 16$

content more representative of a typical day in 2012, and so I found *ReadMe's* better performance on this date reassuring. While the results overall are not ideal, measurement errors across dates and sentiment categories are idiosyncratic enough to reasonably assume they will not affect large-N statistical analysis. To this point, a simple linear regression of *ReadMe* on the hand-coded proportions ($N = 12$) yields a positive correlation ($\beta = .60; p = .13; R^2 = .21$). And as Figure C.1 shows, examination of this regression's residual-vs-fitted plot shows (moderately) well-behaved residuals, suggesting a linear relationship:

Figure C.1: Residual-vs-fitted Plot for *ReadMe* and Hand-coded Proportions: Air Pollution Dispute



As long as the classification error (versus what the "true" hand-coded proportions would be if we had coded samples for all dates across 2012's two phases) is non-systematic, and as long as the *ReadMe* estimates can be modeled as a linear function of the hand codings, this source of error will inflate standard errors but not bias the point estimates.

C.2 Chapter Five

For Chapter Five, I coded four samples of 250 posts each taken from key dates within Phases I-III. Hand-picking these dates also gave me directly estimated proportions for them, which provided a basis for comparison with the algorithm results; in other words, I was able to check, and fine-tune *ReadMe's* performance

by measuring its classification error for those dates. After comparing values for N from 4 to 24, I found that 6 features yielded the lowest absolute-valued classification error across all hand-coded dates. Table C.2 shows key date results:

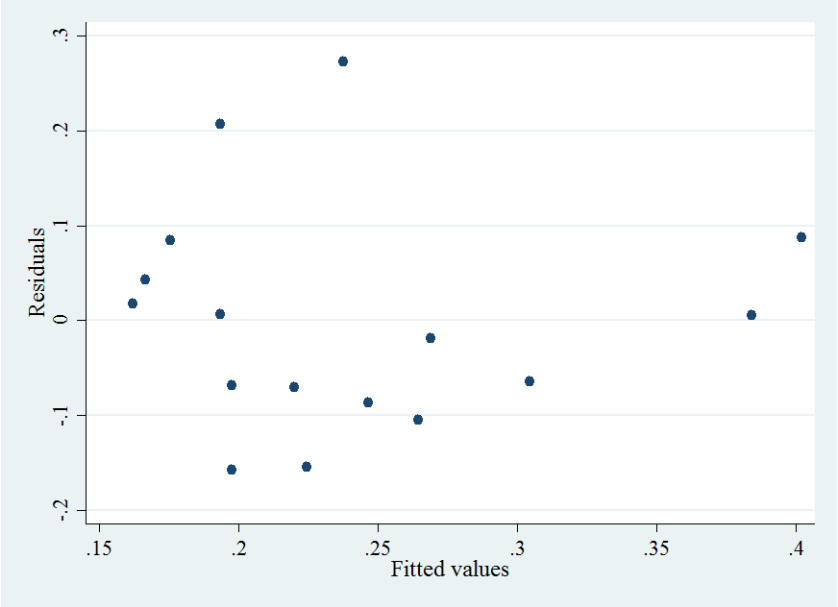
Table C.2: *ReadMe* Versus Hand-Coded Estimates: Bo Xilai Scandal

Date	Measure	Supporters	Questioners	Critics	News
2/8-2/12	*ReadMe*	.13	.16	.51	.16
2/8-2/12	Hand-coded	.10	.21	.19	.25
2/8-2/12	**Abs. Error**	**.03**	**.05**	**.32**	**.09**
3/14-3/16	*ReadMe*	.04	.40	.15	.39
3/14-3/16	Hand-coded	.10	.09	.15	.52
3/14-3/16	**Abs. Error**	**.06**	**.39**	**.00**	**.13**
4/11	*ReadMe*	.26	.25	.18	.24
4/11	Hand-coded	.05	.26	.02	.34
4/11	**Abs. Error**	**.21**	**.01**	**.16**	**.10**
9/28-9/29	*ReadMe*	.21	.20	.07	.49
9/28-9/29	Hand-coded	.03	.09	.16	.56
9/28-9/29	**Abs. Error**	**.18**	**.11**	**.09**	**.07**

$Num.features = 6$

Errors vary widely across dates and categories, with three outliers with absolute valued error greater than 0.2: *Critics* on February 8-12 (.32); *Questioners* on March 14-16 (.39); and *Supporters* on April 11 (.21). While these results are not good, the truly large misses are few enough and idiosyncratic enough to reasonably assume they will not affect large-N statistical analysis. Indeed, a simple linear regression of *ReadMe* on hand-coded proportions ($N = 16$) yields a positive correlation ($\beta = .44; p = .04; R^2 = .26$). And as Figure C.2 shows, examination of this regression's residual-vs-fitted plot shows (moderately) well-behaved residuals, suggesting a linear relationship:

Figure C.2: Residual-vs-fitted Plot for *ReadMe* and Hand-coded Proportions: Bo Xilai Scandal



As in Chapter Four, non-systematic *ReadMe* error will inflate standard errors but not bias point estimates.

C.3 Chapter Six

For Chapter Six, my co-author, our undergraduate assistant, and I hand-coded a total of 479 posts taken from the selected date range. Similar to Chapters Four and Five, I then undertook several trial runs to determine the optimal number of features parameter for *ReadMe* by comparing *ReadMe* estimates with the hand-coded proportions, eventually settling on 8 features. Table C.3 shows these results:

As in previous chapters, the results do show frequent and fairly substantial error. Yet just as before, this does not matter so long as the error is non-systematic

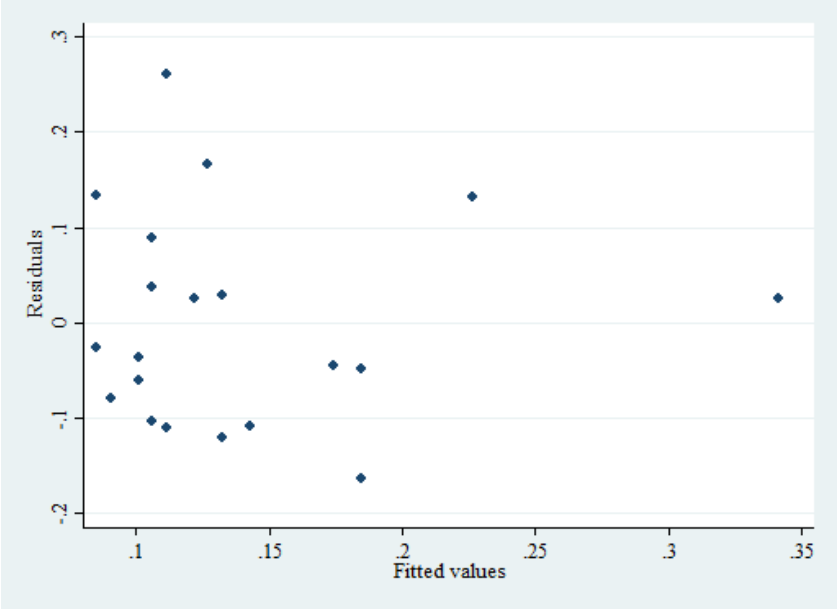
Table C.3: *ReadMe* Versus Hand-Coded Estimates: Diaoyu Dispute

Date	Measure	Moderate	Patriotic	Antijap	Action	Antigov
8/16	*ReadMe*	.003	.034	.065	.161	.358
8/16	Hand-coded	.06	.13	.05	.11	.29
8/16	**Abs. Error**	**.057**	**.096**	**.015**	**.051**	**.068**
8/18	*ReadMe*	.059	.144	.219	.021	.367
8/18	Hand-coded	.02	.06	.02	.21	.51
8/18	**Abs. Error**	**.039**	**.084**	**.199**	**.189**	**.143**
9/11	*ReadMe*	.011	.137	.373	.012	.002
9/11	Hand-coded	.03	.21	.07	.11	.07
9/11	**Abs. Error**	**.019**	**.073**	**.303**	**.098**	**.068**
9/15	*ReadMe*	.129	.148	.04	.196	.293
9/15	Hand-coded	.19	.09	.05	.06	.10
9/15	**Abs. Error**	**.061**	**.058**	**.01**	**.136**	**.193**

$Num.features = 8$

and the *ReadMe* estimates are linearly related to the hand-coded measures. This is the case here as well, with a regression of *ReadMe* on hand-coded proportions ($N = 20$) yielding a positive correlation ($\beta = .52; p = .03; R^2 = .23$). And Figure C.3 shows well-behaved residuals:

Figure C.3: Residual-vs-fitted Plot for *ReadMe* and Hand-coded Proportions: Diaoyu Dispute



When viewing the *ReadMe* results across Chapters 4-6, the degree of error clearly needs reducing. But the above tables and figures suggest that such error is sufficiently non-systematic, and the 'signal' component of the estimates correlated enough with the hand-coded proportions for *ReadMe* results to be used in statistical modeling without being unduly concerned about introducing bias.

BIBLIOGRAPHY

[1] Allison, Graham T. "Conceptual Models and the Cuban Missile Crisis." *American Political Science Review*, vol. 63, no. 03, 1969, pp. 689–718., doi:10.1017/s000305540025853x.

[2] Anderson, Benedict R. O'G. (Benedict Richard O'Gorman). *Imagined Communities: Reflections on the Origin and Spread of Nationalism*. Verso, 1991.

[3] Bamman, David, et al. "Censorship and Deletion Practices in Chinese Social Media." *First Monday*, vol. 17, no. 3, Feb. 2012, doi:10.5210/fm.v17i3.3943.

[4] Bandurski, David. "Innovation, so the Party Can Shine." *China Media Project*, University of Hong Kong, Sept. 2016, cmp.hku.hk/2016/09/02/39893/. Accessed Sept. 2016.

[5] Bennett, Philip, and Moises Naim. "21st Century Censorship: Governments around the World Are Using Stealthy Strategies to Manipulate the Media." *Columbia Journalism Review*, 2015.

[6] Blei, David, et al. "Latent Direchlet Allocation." *Journal of Machine Learning Research*, vol. 3, pp. 993–1022.

[7] Brady, Anne-Marie. *Marketing Dictatorship: Propaganda and Thought Work in Contemporary China*. Lanham, MD, Rowman & Littlefield, 2008.

[8] Brady, Henry E., and David Collier. *Rethinking Social Inquiry: Diverse Tools, Shared Standards*. Lanham, MD, Rowman & Littlefield Publishers, 2004.

[9] Broadhurst, Roderic, and Peng Wang. "After the Bo Xilai Trial: Does Corruption Threaten China's Future?" *Survival*, vol. 56, no. 3, Apr. 2014, pp. 157–178., doi:10.1080/00396338.2014.920148.

[10] Brodsgaard, Kjeld Erik. *Hainan: State, Society and Business in a Chinese Province*. London, Routledge, 2008.

[11] Cairns, Christopher, and Allen Carlson. "Real-World Islands in a Social Media Sea: Nationalism and Censorship on Weibo during the 2012 Diaoyu/Senkaku Crisis." *The China Quarterly*, vol. 225, 2016, pp. 23–49., doi:10.1017/s0305741015001708.

[12] Cairns, Christopher, and Elizabeth Plantan. "Why Autocrats Sometimes Relax Online Censorship of Sensitive Issues: A Case Study of Microblog Discussion of Air Pollution in China." Working Paper, October 2016, http://www.chrismcairns.com

[13] Cairns, Christopher, and Elizabeth Plantan. "Hazy Messaging: Framing Air Pollution on Chinese Social Media." Working Paper, October 2016, http://www.chrismcairns.com

[14] Callahan, W. A. "The Cartography of National Humiliation and the Emergence of China's Geobody." *Public Culture*, vol. 21, no. 1, Jan. 2009, pp. 141–173., doi:10.1215/08992363-2008-024.

[15] Carlson, Allen. "A Flawed Perspective: the Limitations Inherent within the Study of Chinese Nationalism." *Nations and Nationalism*, vol. 15, no. 1, 2009, pp. 20–35., doi:10.1111/j.1469-8129.2009.00376.x.

[16] Chan, Chak K., and Xiaohong Yao. "Air Pollution in Mega Cities in China." *Atmospheric Environment*, vol. 42, no. 1, 2008, pp. 1–42., doi:10.1016/j.atmosenv.2007.09.003.

[17] Chen, Chih-Jou. "Growing Social Unrest and Emergent Protest Groups in China." *Rise of China: Beijing's Strategies and Implications for the Asia-Pacific*, edited by H.M. Hsiao and C. Lin, Routledge, New York, pp. 87–106.

[18] Chen, Jidong, and Yiqing Xu. "Why Do Authoritarian Regimes Allow Citizens to Voice Opinions Publicly?" *Journal of Politics*, forthcoming 2016. SSRN: https://ssrn.com/abstract=2318051 or http://dx.doi.org/10.2139/ssrn.2318051

[19] Chen, Lidan. "Open Information System and Crisis Communication in China." *Chinese Journal of Communication*, vol. 1, no. 1, 2008, pp. 38–54., doi:10.1080/17544750701861905.

[20] Chen, Ni. "Institutionalizing Public Relations: A Case Study of Chinese Government Crisis Communication on the 2008 Sichuan Earthquake." *Public Relations Review*, vol. 35, no. 3, 2009, pp. 187–198., doi:10.1016/j.pubrev.2009.05.010.

[21] Cheng, Yinghong. "From Campus Racism to Cyber Racism: Discourse of Race and Chinese Nationalism." *The China Quarterly*, vol. 207, 2011, pp. 561–579., doi:10.1017/s0305741011000658.

[22] Creemers, Rogier. "Cyber China: Upgrading Propaganda, Public Opinion Work and Social Management for the Twenty-First Century." *Journal of Contemporary China*, vol. 26, no. 103, May 2016, pp. 85–100., doi:10.1080/10670564.2016.1206281.

[23] Cui, Di, and Fang Wu. "Moral Goodness and Social Orderliness: An Analysis of the Official Media Discourse about Internet Governance in China." *Telecommunications Policy*, no. 40, 2016, pp. 265–276., doi:http://doi.org/10.1016/j.telpol.2015.11.010.

[24] Fewsmith, Joseph. "China's 18th Party Congress: What's at Stake?" *Current History*, no. 111, Sept. 2012, pp. 203–208.

[25] Converse, Philip E. *The Nature of Belief Systems in Mass Publics*. Ann Arbor, Survey Research Center, Univ. of Michigan, 1962.

[26] Dimitrov, Martin K. "Internal Government Assessments of the Quality of Governance in China." *Studies in Comparative International Development*, vol. 50, no. 1, 2014, pp. 50–72., doi:10.1007/s12116-014-9170-2.

[27] Dimitrov, Martin K. "The Resilient Authoritarians." *Current History*, Jan. 2008, pp. 24–29.

[28] Dimitrov, Martin K. "Tracking Public Opinion Under Authoritarianism." *Russian History*, vol. 41, no. 3, 2014, pp. 329–353., doi:10.1163/18763316-04103003.

[29] Dimitrov, Martin K. "What the Party Wanted to Know." *East European Politics and Societies*, vol. 28, no. 2, 2014, pp. 271–295., doi:10.1177/0888325413506933.

[30] Egorov, Georgy, et al. "Why Resource-Poor Dictators Allow Freer Media: A Theory and Evidence from Panel Data." *American Political Science Review*, vol. 103, no. 04, 2009, pp. 645–668., doi:10.1017/s0003055409990219.

[31] Esarey, Ashley, and Xiao Qiang. "Digital Communication and Political Change in China." *International Journal of Communication*, vol. 5, 2011, pp. 298–319.

[32] Esarey, Ashley. "Understanding Chinese Regime Censorship and Preferences." Presented at the American Political Science Association Annual Meeting, Chicago, IL, September 1, 2013.

[33] Esarey, Ashley. "Winning Hearts and Minds? Cadres as Microbloggers in China." *Journal of Current Chinese Affairs*, vol. 2, 2015, pp. 69–103.

[34] Fewsmith, Joseph. "Promoting the Scientific Development Concept." *China Leadership Monitor*, no. 11, 2004, pp. 1–10.

[35] Freedom House. Freedom on the Net: 2009-16. freedomhouse.org/report/freedom-net/.

[36] Fu, King-Wa, and Michael Chau. "Reality Check for the Chinese Microblog Space: A Random Sampling Approach." *PLoS ONE* 8(3): e58356, doi:10.1371/journal.pone.0058356.

[37] Fu, King-Wa, et al. "Assessing Censorship on Microblogs in China: Discriminatory Keyword Analysis and the Real-Name Registration Policy." *IEEE Internet Computing*, vol. 17, no. 3, 2013, pp. 42–50., doi:10.1109/mic.2013.28.

[38] Gandhi, Jennifer, and Adam Przeworski. "Cooperation, Cooptation, And Rebellion Under Dictatorships." *Economics and Politics*, vol. 18, no. 1, 2006, pp. 1–26., doi:10.1111/j.1468-0343.2006.00160.x.

[39] Gandhi, Jennifer. *Political Institutions under Dictatorship*. Cambridge, Cambridge University Press, 2008.

[40] Gasiorowski, Mark J. "Economic Crisis and Political Regime Change: An Event History Analysis." *American Political Science Review*, vol. 89, no. 04, 1995, pp. 882–897., doi:10.2307/2082515.

[41] Geddes, Barbara, and John Zaller. "Sources of Popular Support for Authoritarian Regimes." *American Journal of Political Science*, vol. 33, no. 2, 1989, p. 319., doi:10.2307/2111150.

[42] Gehlbach, Scott, and Konstantin Sonin. "Government Control of the Media." *SSRN Electronic Journal*, doi:10.2139/ssrn.1315882.

[43] Granger, C. W. J. "Investigating Causal Relations by Econometric Models and Cross-Spectral Methods." *Econometrica*, vol. 37, no. 3, 1969, p. 424., doi:10.2307/1912791.

[44] Granovetter, Mark. "Threshold Models of Collective Behavior." *American Journal of Sociology*, vol. 83, no. 6, 1978, pp. 1420–1443., doi:10.1086/226707.

[45] Gries, Peter Hays. *China's New Nationalism: Pride, Politics, and Diplomacy.* Berkeley, University of California Press, 2004.

[46] Gries, Peter Hays, et al. "Patriotism, Nationalism and China's US Policy: Structures and Consequences of Chinese National Identity." *The China Quarterly*, vol. 205, 2011, pp. 1–17., doi:10.1017/s0305741010001360.

[47] Gueorguiev, Dimitar D., and Paul J. Schuler. "Keeping Your Head Down: Public Profiles And Promotion Under Autocracy." *Journal of East Asian Studies*, vol. 16, no. 01, 2016, pp. 87–116., doi:10.1017/jea.2015.1.

[48] Gunitsky, Seva. "Corrupting the Cyber-Commons: Social Media as a Tool of Autocratic Stability." *Perspectives on Politics*, vol. 13, no. 01, 2015, pp. 42–54., doi:10.1017/s1537592714003120.

[49] Guriev, Sergei, and Daniel Treisman. "How Modern Dictators Survive: An Informational Theory of the New Authoritarianism." NBER Working Paper, National Bureau of Economic Research, Apr. 2015, www.nber.org/papers/w21136. Accessed 9 Mar. 2017.

[50] Han, Rongbin. "Manufacturing Consent in Cyberspace: China's 'Fifty-Cent Army'." *Journal of Current Chinese Affairs*, vol. 2, 2015, pp. 105–134.

[51] Hanley, J. A. "If Nothing Goes Wrong, Is Everything All Right? Interpreting Zero Numerators." *JAMA: The Journal of the American Medical Association*, vol. 249, no. 13, Jan. 1983, pp. 1743–1745., doi:10.1001/jama.249.13.1743.

[52] Hassid, Jonathan, and Wanning Sun. "Stability Maintenance and Chinese Media: Beyond Political Communication?" *Journal of Current Chinese Affairs*, Vol. 44, no. 2, 2015, pp. 3–15.

[53] Hassid, Jonathan. "Safety Valve or Pressure Cooker? Blogs in Chinese Political Life." *Journal of Communication*, vol. 62, no. 2, 2012, pp. 212–230., doi:10.1111/j.1460-2466.2012.01634.x.

[54] Hay, Colin. "Crisis and the Structural Transformation of the State: Interrogating the Process of Change." *The British Journal of Politics and International Relations*, vol. 1, no. 3, 1999, pp. 317–344., doi:10.1111/1467-856x.00018.

[55] He, Baogang, and Mark E. Warren. "Authoritarian Deliberation: The Deliberative Turn in Chinese Political Development." *Perspectives on Politics*, vol. 9, no. 02, 2011, pp. 269–289., doi:10.1017/s1537592711000892.

[56] He, Yinan. "History, Chinese Nationalism, and the Emerging Sino-Japanese Conflict." *Journal of Contemporary China*, vol. 16, no. 50, 2007, pp. 1–24.

[57] Hildebrandt, Timothy, and Jennifer Turner. "Green Activism? Reassessing the Role of Environmental NGOs in China." *State and Society Responses to Social Welfare Needs in China: Serving the People*, edited by J Schwartz and S Shieh, Routledge, New York, 2009, pp. 89–110.

[58] Ho, Peter, and Richard Louis Edmonds. *China's Embedded Activism: Opportunities and Constraints of a Social Movement*. London, Routledge, 2008.

[59] Ho, Peter. "Greening Without Conflict? Environmentalism, NGOs and Civil Society in China." *Development and Change*, vol. 32, no. 5, 2001, pp. 893–921., doi:10.1111/1467-7660.00231.

[60] Hoffmann, Robert, and Jeremy Larner. "The Demography of Chinese Nationalism: A Field-Experimental Approach." *The China Quarterly*, vol. 213, 2013, pp. 189–204., doi:10.1017/s0305741013000271.

[61] Hooren, Franca Van, et al. "The Shock Routine: Economic Crisis and the Nature of Social Policy Responses." *Journal of European Public Policy*, vol. 21, no. 4, 2014, pp. 605–623., doi:10.1080/13501763.2014.887757.

[62] Hopkins, Daniel J., and Gary King. "A Method of Automated Nonparametric Content Analysis for Social Science." *American Journal of Political Science*, vol. 54, no. 1, 2010, pp. 229–247., doi:10.1111/j.1540-5907.2009.00428.x.

[63] Howard, Philip N., and Muzammil M. Hussain. "The Role of Digital Media." *Journal of Democracy*, vol. 22, no. 3, 2011, pp. 35–48., doi:10.1353/jod.2011.0041.

[64] Hu, Yong. "Three Features of China's Internet Regulation." *Chinese Law and Government*, vol. 48, no. 1, 2016, pp. 1–5.

[65] Hughes, Christopher R. *Chinese Nationalism in the Global Era*. London, Routledge, 2006. Hussain, Muzammil M., and Philip N. Howard. "What Best Explains Successful Protest Cascades? ICTs and the Fuzzy Causes of the Arab Spring." *International Studies Review*, vol. 15, no. 1, 2013, pp. 48–66., doi:10.1111/misr.12020.

[66] Iyengar, Shanto, and Donald R. Kinder. *News That Matters: Television and American Opinion*. Chicago (Ill.), The University of Chicago Press, 2010.

[67] Johnston, Alastair Iain. "Thinking about Strategic Culture." *International Security*, vol. 19, no. 4, 1995, p. 32., doi:10.2307/2539119.

[68] Johnston, Alastair Iain, and Daniela Stockmann. "Chinese Attitudes toward the United States and Americans." *Anti-Americanisms in World Politics*, edited by Peter Katzenstein and Robert Keohane, Cornell Univ. Press, Ithaca, NY, 2007.

[69] Ke Li, Angela. "Towards a More Proactive Method: Regulating Public Opinion on Chinese Microblogs under Xi's New Leadership." *China Perspectives*, no. 4, 2015, pp. 15–23.

[70] Kennedy, John James. "Maintaining Popular Support for the Chinese Communist Party: The Influence of Education and the State-Controlled Media." *Political Studies*, vol. 57, no. 3, 2009, pp. 517–536., doi:10.1111/j.1467-9248.2008.00740.x.

[71] Keohane, Robert O., and Joseph S. Nye. *Power and Interdependence*. Boston, Little, Brown, 1977.

[72] King, G., et al. "Reverse-Engineering Censorship in China: Randomized Experimentation and Participant Observation." *Science*, vol. 345, no. 6199, 2014, pp. 1251722–1251722., doi:10.1126/science.1251722.

[73] King, Gary, et al. *Designing Social Inquiry Scientific Inference in Qualitative Research*. Princeton, NJ, Princeton Univ. Press, 1994.

[74] King, Gary, et al. "How Censorship in China Allows Government Criticism but Silences Collective Expression." *American Political Science Review*, vol. 107, no. 02, 2013, pp. 326–343., doi:10.1017/s0003055413000014.

[75] King, Gary, et al. "How the Chinese Government Fabricates Social Media Posts for Strategic Distraction, Not Engaged Argument." *American Political Science Review*, 2017.

[76] Kuran, Timur. "Now out of Never: The Element of Surprise in the East European Revolution of 1989." *World Politics*, vol. 44, no. 01, 1991, pp. 7–48., doi:10.2307/2010422.

[77] Kuran, Timur. *Private Truths, Public Lies: the Social Consequences of Preference Falsification*. Cambridge (Mass.), Harvard University Press, 1995.

[78] Kuran, Timur. "Sparks and Prairie Fires: A Theory of Unanticipated Political Revolution." *Public Choice*, vol. 61, no. 1, 1989, pp. 41–74., doi:10.1007/bf00116762.

[79] Lam, Oiwan. "China Beefs up '50 Cent' Army of Paid Internet Propagandists." *Global Voices*, Oct. 2013.

[80] Lang, Graeme, and Ying Xu. "Anti-Incinerator Campaigns and the Evolution of Protest Politics in China." *Environmental Politics*, vol. 22, no. 5, 2013, pp. 832–848., doi:10.1080/09644016.2013.765684.

[81] Lei, Ya-Wen. "The Political Consequences of the Rise of the Internet: Political Beliefs and Practices of Chinese Netizens." *Political Communication*, vol. 28, no. 3, 2011, pp. 291–322., doi:10.1080/10584609.2011.572449.

[82] Lewis, Orion A. "Net Inclusion: New Media's Impact on Deliberative Politics in China." *Journal of Contemporary Asia*, vol. 43, no. 4, 2013, pp. 678–708., doi:10.1080/00472336.2013.769387.

[83] Lewis-Beck, Michael S., et al. "A Chinese Popularity Function." *Political Research Quarterly*, vol. 67, no. 1, 2014, pp. 16–25., doi:10.1177/1065912913486196.

[84] Li, Jinshan. *Bureaucratic Restructure in Reforming China: a Redistribution of Political Power*. Singapore, World Scientific, 1998.

[85] Lieberthal, Kenneth, and Michel Oksenberg. *Policy Making in China Leaders, Structures, and Processes*. Princeton, N. J., Princeton University, 1988.

[86] Lieberthal, Kenneth G., and David M. Lampton. *Bureaucracy, Politics and Decision Making in Post-Mao China*. Berkeley (Calif.), University of California Press, 1992.

[87] Lieberthal, Kenneth G. *Governing China from Revolution through Reform*. New York, W.W. Norton, 1995.

[88] Liebman, Benjamin L. "The Media and the Courts: Towards Competitive Supervision?" *The China Quarterly*, vol. 208, 2011, pp. 833–850., doi:10.1017/s0305741011001020.

[89] Link, Perry, and Xiao Qiang. "From 'Fart People' to Citizens." *Journal of Democracy*, vol. 24, no. 1, 2013, pp. 79–85., doi:10.1353/jod.2013.0014.

[90] Little, Andrew T. "Communication Technology and Protest." *The Journal of Politics*, vol. 78, no. 1, 2016, pp. 152–166., doi:10.1086/683187.

[91] Lohmann, Susanne. "Collective Action Cascades: An Informational Rationale for the Power in Numbers." *Journal of Economic Surveys*, vol. 14, no. 5, 2000, pp. 655–684., doi:10.1111/1467-6419.00128.

[92] Lohmann, Susanne. "The Dynamics of Informational Cascades: The Monday Demonstrations in Leipzig, East Germany, 1989-91." *World Politics*, vol. 47, no. 01, 1994, pp. 42–101., doi:10.2307/2950679.

[93] Lorentzen, Peter. "China's Strategic Censorship." *American Journal of Political Science*, vol. 58, no. 2, Aug. 2013, pp. 402–414., doi:10.1111/ajps.12065.

[94] Lu, Xiaobo. "Social Policy and Regime Legitimacy: The Effects of Education Reform in China." *American Political Science Review*, vol. 108, no. 02, 2014, pp. 423–437., doi:10.1017/s0003055414000124.

[95] Lucas, Christopher, et al. "Computer-Assisted Text Analysis for Comparative Politics." *Political Analysis*, vol. 23, no. 02, 2015, pp. 254–277., doi:10.1093/pan/mpu019.

[96] Lynch, Daniel C. *After the Propaganda State: Media, Politics, and "Thought Work" in Reformed China*. Stanford, CA, Stanford Univ. Press, 1999.

[97] Lynch, Marc. "After Egypt: The Limits and Promise of Online Challenges to the Authoritarian Arab State." *Perspectives on Politics*, vol. 9, no. 02, 2011, pp. 301–310., doi:10.1017/s1537592711000910.

[98] Mackie, John L. *The Cement of the Universe: a Study of Causation*. Oxford, Clarendon Pr., 1974.

[99] MacKinnon, Rebecca. *Consent of the Networked: the Worldwide Struggle for Internet Freedom*. New York, Basic Books, 2012.

[100] Mccombs, Maxwell E., and Donald L. Shaw. "The Agenda-Setting Function of Mass Media." *Public Opinion Quarterly*, vol. 36, no. 2, 1972, p. 176., doi:10.1086/267990.

[101] Meng, Tianguang, et al. "Conditional Receptivity to Citizen Participation." *Comparative Political Studies*, vol. 50, no. 4, 2017, pp. 399–433., doi:10.1177/0010414014556212.

[102] Mertha, Andrew. *China's Water Warriors: Citizen Action and Policy Change*. Ithaca, NY, Cornell University Press, 2008.

[103] Mertha, Andrew. *The Politics of Piracy: Intellectual Property in Contemporary China*. Ithaca, Cornell University Press, 2005.

[104] Mesquita, Bruce Bueno de. *Logic of Political Survival*. MIT Press, 2003.

[105] Miller, Alice. "The Trouble with Factions." *China Leadership Monitor*, vol. 46, 2015, pp. 1–12.

[106] Miller, Blake, and Mary Gallagher. "Astroturfing in China: Three Case Studies." Research Report, 2017. http://www.blakeapm.com/research/. Accessed 9 Mar. 2017.

[107] Morozov, Evgeny. *The Net Delusion: the Dark Side of Internet Freedom*. New York, Public Affairs, 2011.

[108] Nathan, Andrew J. (Andrew James). "Authoritarian Resilience." *Journal of Democracy*, vol. 14, no. 1, 2003, pp. 6–17., doi:10.1353/jod.2003.0019.

[109] Naughton, Barry. "Reform, Retreat and Renewal: How Economic Policy Fits into the Political System." *Issues and Studies*, vol. 51, no. 1, 2015, pp. 23–54.

[110] Ng, Jason and Pierre Landry. "The Political Hierarchy of Censorship: An Analysis of Keyword Blocking of CCP Officials' Names on Sina Weibo Before and After the 2012 National Congress (S)election." Eleventh Chinese Internet Research Conference, 2013, Forthcoming. SSRN: https://ssrn.com/abstract=2267367

[111] Ng, Jason. "Justifying Censorship: Using WeChat Public Posts to Examine the Chinese Initiative to Curtail Online Falsehoods (Anti-Rumor Campaign)." Presented at the China Internet Research Conference, Edmonton, Alberta, May 27-28, 2015.

[112] Ng, Jason. "Tracing the Path of a Censored Weibo Post and Compiling Keywords That Trigger Automatic Review." The Citizen Lab, 1 Mar. 2015, citizenlab.org/2014/11/tracing-path-censored-weibo-post-compiling-keywords-trigger-automatic-review/. Accessed 9 Mar. 2017.

[113] Noesselt, Nele. "Microblogs and the Adaptation of the Chinese Party-State's

Governance Strategy." *Governance*, vol. 27, no. 3, Jan. 2013, pp. 449–468., doi:10.1111/gove.12045.

[114] Oksenberg, Michel. "China's Confident Nationalism." *Foreign Affairs*, vol. 65, no. 3, 1986, pp. 501–523., doi:10.2307/20043078.

[115] Patel, David Siddharta. "Roundabouts and Revolutions: Public Squares, Coordination, and the Diffusion of the Arab Uprisings." Working Paper, 2013.

[116] Reilly, James. *Strong Society, Smart State: the Rise of Public Opinion in China's Japan Policy*. New York, Columbia University Press, 2011.

[117] Ren, Xianliang. "*Tongchou Liangge Yulunchang, Ningju Shehui Zhengnengliang*" ["Comprehensively Plan Both Public Opinion Fields, Concentrate Social Positive Energy"]. *Hongqi Wengao* [*Red Flag Manuscripts*] no. 7, 2013.

[118] Roberts, Margaret, et al. "Stm: R Package for Structural Topic Models." *Journal of Statistical Software*, forthcoming.

[119] Roberts, Margaret E. "Fear, Friction and Flooding: Methods of Online Information Control." Doctoral dissertation, *Harvard University*, 2014.

[120] Roberts, Margaret E. "Experiencing Censorship Emboldens Internet Users and Decreases Government Support in China." Working Paper, 2015. http://margaretroberts.net/wp-content/uploads/2015/07/fear.pdf

[121] Rozman, Gilbert. "Chinese National Identity and East Asian National Identity Gaps." *National Identities and Bilateral Relations: Widening Gaps and Chinese Demonization of the United States*, edited by Gilbert Rozman, Stanford University Press, Stanford, CA, 2013, pp. 203–233.

[122] Schelling, Thomas C. *The Strategy of Conflict*. Cambridge, Harvard University Press, 1960.

[123] Shadmehr, Mehdi, and Dan Bernhardt. "State Censorship." *American Economic Journal: Microeconomics*, vol. 7, no. 2, 2015, pp. 280–307., doi:10.1257/mic.20130221.

[124] Shambaugh, David. "China's Propaganda System: Institutions, Processes and Efficacy." *The China Journal*, vol. 57, 2007, pp. 25–58., doi:10.1086/tcj.57.20066240.

[125] Shen, Simon. *Redefining Nationalism in Modern China: Sino -American Relations and the Emergence of Chinese Public Opinion in the 21st Century.* Palgrave Macmillan, 2007.

[126] Shirk, Susan L., editor. *Changing Media, Changing China.* New York, Oxford University Press, 2011.

[127] Sima, Yangzi. "Grassroots Environmental Activism and the Internet: Constructing a Green Public Sphere in China." *Asian Studies Review*, vol. 35, no. 4, 2011, pp. 477–497., doi:10.1080/10357823.2011.628007.

[128] Skinner, G. William. "Marketing and Social Structure in Rural China: Part I." *The Journal of Asian Studies*, vol. 24, no. 1, 1964, p. 3., doi:10.2307/2050412.

[129] Stockmann, Daniela, and Mary E. Gallagher. "Remote Control: How the Media Sustain Authoritarian Rule in China." *Comparative Political Studies*, vol. 44, no. 4, 2011, pp. 436–467., doi:10.1177/0010414010394773.

[130] Stockmann, Daniela and Ting Luo. "Which Social Media Facilitate Online Public Opinion in China?" July 2015. Available at SSRN: https://ssrn.com/abstract=2663018 or http://dx.doi.org/10.2139/ssrn.2663018

[131] Stockmann, Daniela. *Media Commercialization and Authoritarian Rule in China.* Cambridge, UK, Cambridge Univ. Press, 2013.

[132] Stockmann, Daniela. "Propaganda for Sale: The Impact of Newspaper Commercialization on News Content and Public Opinion in China." Doctoral dissertation, *University of Michigan*, 2007.

[133] Stockmann, Daniela. "Who Believes Propaganda? Media Effects during the Anti-Japanese Protests in Beijing." *The China Quarterly*, vol. 202, 2010, pp. 269–289., doi:10.1017/s0305741010000238.

[134] Tai, Qiuqing. "China's Media Censorship: A Dynamic and Diversified Regime." *Journal of East Asian Studies*, vol. 14, no. 02, 2014, pp. 185–210., doi:10.1017/s1598240800008900.

[135] Tang, Wenfang. *Public Opinion and Political Change in China.* Stanford, Stanford University Press, 2005.

[136] Teets, Jessica C. "Let Many Civil Societies Bloom: The Rise of Consultative Authoritarianism in China." *The China Quarterly*, vol. 213, 2013, pp. 19–38., doi:10.1017/s0305741012001269.

[137] Tilt, Bryan, and Qing Xiao. "Media Coverage of Environmental Pollution in the People's Republic of China: Responsibility, Cover-up and State Control." *Media, Culture & Society*, vol. 32, no. 2, 2010, pp. 225–245., doi:10.1177/0163443709355608.

[138] Tong, Yanqi, and Shaohua Lei. "War of Position and Microblogging in China." *Journal of Contemporary China*, vol. 22, no. 80, 2013, pp. 292–311., doi:10.1080/10670564.2012.734084.

[139] Truex, Rory. "Consultative Authoritarianism and Its Limits." *Comparative Political Studies*, vol. 50, no. 3, 2017, pp. 329–361., doi:10.1177/0010414014534196.

[140] Volkov, Denis, and Stepan Goncharov. Rossiiskii Media-Landshaft: Televidenie, Pressa, Internet [The Russian Media Landscape: Television, Press, Internet]. *Levada Center*, 17 June 2014, www.levada.ru/17-06-2014/rossiiskii-media-landshaft-televidenie-pressa-internet. Accessed 9 Mar. 2017.

[141] Wang, Zhengxu. "Hu Jintao's Power Consolidation: Groups, Institutions, and Power Balance in China's Elite Politics." *Issues and Studies*, vol. 42, no. 4, 2006, pp. 97–136.

[142] Weeks, Jessica L. "Autocratic Audience Costs: Regime Type and Signaling Resolve." *International Organization*, vol. 62, no. 01, 2008, doi:10.1017/s0020818308080028.

[143] Weiss, Jessica Chen. "Authoritarian Signaling, Mass Audiences, and Nationalist Protest in China." *International Organization*, vol. 67, no. 01, 2013, pp. 1–35., doi:10.1017/s0020818312000380.

[144] Weiss, Jessica Chen. *Powerful Patriots: Nationalist Protest in China's Foreign Relations*. by Jessica Chen Weiss. New York, Oxford University Press, 2014.

[145] Whiting, Allen S. "Assertive Nationalism in Chinese Foreign Policy." *Asian Survey*, vol. 23, no. 8, 1983, pp. 913–933., doi:10.2307/2644264.

[146] Whitten-Woodring, Jenifer, and Patrick James. "Fourth Estate or Mouth-

piece? A Formal Model of Media, Protest, and Government Repression." *Political Communication*, vol. 29, no. 2, 2012, pp. 113–136., doi:10.1080/10584609.2012.671232.

[147] Willis, Michael J. *Politics and Power in the Maghreb: Algeria, Tunisia and Morocco from Independence to the Arab Spring*. Oxford Univ. Press, 2012.

[148] Wintrobe, Ronald. *The Political Economy of Dictatorship*. Cambridge, Cambridge University Press, 1998.

[149] *World Factbook*. Central Intelligence Agency, 1 Apr. 2016, www.cia.gov/library/publications/the-world-factbook/. Accessed 8 Mar. 2017.

[150] World Values Survey Wave 5 2005-2008 OFFICIAL AGGREGATE v.20140429. World Values Survey Association (www.worldvaluessurvey.org). Aggregate File Producer: Asep/JDS, Madrid SPAIN.

[151] World Values Survey Wave 6 2010-2014 OFFICIAL AGGREGATE v.20150418. World Values Survey Association (www.worldvaluessurvey.org). Aggregate File Producer: Asep/JDS, Madrid SPAIN.

[152] Xiao, Qiang. "The Battle for the Chinese Internet." *Journal of Democracy*, vol. 22, no. 2, 2011, pp. 47–61., doi:10.1353/jod.2011.0020.

[153] Yang, G., and C. Calhoun. "Media, Civil Society, and the Rise of a Green Public Sphere in China." *China Information*, vol. 21, no. 2, Jan. 2007, pp. 211–236., doi:10.1177/0920203x07079644.

[154] Yang, Guobin. "The Return of Ideology and the Future of Chinese Internet Policy." *Critical Studies in Media Communication*, vol. 31, no. 2, 2014, pp. 109–113., doi:10.1080/15295036.2014.913803.

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

[156] Yuen, Samson. "Disciplining the Party: Xi Jinping's Anti-Corruption Campaign and Its Limits." *China Perspectives*, no. 3, 2014, pp. 41–47.

[157] Zaller, John R. *The Nature and Origins of Mass Opinion*. Cambridge University Press, 1992.

[158] Zhang, Xiaoling. "From Totalitarianism to Hegemony: the Reconfiguration of the Party-State and the Transformation of Chinese Communication." *Journal of Contemporary China*, vol. 20, no. 68, Sept. 2010, pp. 103–115., doi:10.1080/10670564.2011.520850.

[159] Zhao, Suisheng. *A Nation-State by Construction: Dynamics of Modern Chinese Nationalism*. Stanford, CA, Stanford Univ. Press, 2004.

[160] Zhao, Yuezhi. "From Commercialization to Conglomeration: the Transformation of the Chinese Press within the Orbit of the Party State." *Journal of Communication*, vol. 50, no. 2, 2000, pp. 3–26., doi:10.1093/joc/50.2.3.

[161] Zhao, Yuezhi. *Media, Market, and Democracy in China: between the Party Line and the Bottom Line*. Urbana, Univ. of Illinois Press, 1998.

[162] Zheng, Yongnian. *Technological Empowerment: the Internet, State, and Society in China*. Stanford, Calif, Stanford University Press, 2007.

[163] Zhou, Yongming. *Historicizing Online Politics Telegraphy, the Internet, and Political Participation in China*. Stanford, Stanford University Press, 2006.

[164] Zhu, Jiangnan, et al. "When Grapevine News Meets Mass Media: Different Information Sources and Popular Perceptions of Government Corruption in Mainland China." *Comparative Political Studies*, vol. 46, no. 8, 2012, pp. 920–946., doi:10.1177/0010414012463886.

[165] Zhu, Tao, et al. "The Velocity of Censorship: High-Fidelity Detection of Microblog Post Deletions." [1303.0597] Cornell University Library, 10 July 2013, arxiv.org/abs/1303.0597. Accessed 9 Mar. 2017.