Delegated Dictatorship: Examining the State and Market Forces behind Information Control in China

by

Blake Miller

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Political Science and Scientific Computing)
in The University of Michigan
2018

Doctoral Committee:

Professor Mary Gallagher, Chair
Professor Pauline Jones
Professor Walter Mebane
Professor Nicholas Valentino

Blake Miller

blakeapm@umich.edu

ORCID iD: 0000-0002-4707-0984

ACKNOWLEDGEMENTS

The seed of this dissertation was planted in 2008 when I was fortunate enough to meet Larry Diamond and to take the excellent Chinese Politics class taught by Alice Miller as an undergraduate at Stanford University. I spent the next 3 years as Larry's research assistant, studying internet censorship in China, and became fascinated by the online conflict between state and society that seemed to be pushing China in a more open, liberal, and democratic direction. Larry encouraged me to go to China and conduct fieldwork for an undergraduate thesis on the use of the internet to control dissent. Semi-structured interviews with the signers and drafters of Charter 08 opened my eyes to the real-world consequences of online dissent in authoritarian regimes. This experience provided a strong foundation upon which I built my graduate school research agenda. I am incredibly grateful to Alice Miller and Larry Diamond for their support and encouragement at this early stage.

Going through graduate school and writing a dissertation is an extraordinarily difficult and stressful process. I was very fortunate to have many kind, thoughtful, and supporting friends, family members, mentors, and research assistants who made the process easier, and an incredibly meaningful and rewarding experience.

Thanks to my parents, Blake and Jeryl Miller for your encouragement and support throughout grad school. Thanks as well to my brother Luke Miller and his partner Lisa Quadt, two grad school veterans who are always supportive and thoughtful.

To the friends and fellow graduate students I met on my first day at Michigan, Sasha de Vogel, Zander Furnas, Joseph Klaver, Steven Moore, and Michael

Thompson-Brusstar, thanks for being the most amazing group of friends throughout the entirety of my time at Michigan. You were always supportive when I doubted my work, always listened to me when I was feeling discouraged, and always made sure that I took breaks when I needed them. To Nicole Wu, thanks for being a great friend throughout my time at Michigan, for sending me ALL the cute dog pictures, and for double-checking my attempts to translate Mandarin into English. Thanks to all of the other friends and acquaintances in Ann Arbor who made grad school a pleasant and memorable experience.

Thanks to my dissertation committee Mary Gallagher, Walter Mebane, Pauline Jones, and Nicholas Valentino for their incredible feedback, generosity, and thoughtfulness. Mary, thanks for your honest and frank feedback in our many meetings together, for helping me focus and expand my research agenda when necessary, for fostering a community of grad students studying China, and for taking the time to introduce me to scholars outside of Michigan who have mentored me throughout my time at Michigan. Walter, thank you for encouraging me in my interest in machine learning and natural language processing early on in grad school, for your generosity with time and resources, and for allowing me to use your personal server for data collection, even when it slowed your own work down (sorry about that again). Pauline, thanks for keeping me honest when it came to theory development and for pushing me to be more systematic in my research process. Thank you Nick for helping me with the design of my survey experiments, and for telling me not to worry when I was unnecessarily concerned about outcomes of grants, funding, and the job market.

Thank you to Daniela Stockmann, who hosted me multiple times in Berlin and in Shanghai, and who has become an incredible mentor and friend. Thanks to Yuhua Wang, Iza Ding, and Jeffrey Javed for mentorship and feedback during grad school that undoubtedly has made me a better scholar, and has improved the quality of my work.

Thanks to Yuen Yuen Ang, Christian Davenport, Chris Fariss, Allen Hicken, Brian Min, James Morrow, Ragnhild Nordas, Iain Osgood, and Charles Shipan for their feedback at workshops and meetings at the University of Michigan. Thank you to Margaret Roberts, Haifeng Huang, Pierre Landry, and Bruce Dickson for your feedback on my work.

This dissertation would not have been possible without the excellent research assistantship of Emma, Leon, Shannon, Qiwei, Moira, Zhang, Martin, Chen, Erin, Yingsi, Jack, Ziyi, and Sherry. Thanks for your dedication and hard work that has resulted in the datasets used in this dissertation.

Thanks to Sweetland Writing Center at the University of Michigan for arranging dissertation writing groups and providing excellent writing help. Thanks to Simone Sessolo for reading many drafts of my chapters. Thanks to my dissertation writing group members Angie Baecker, Chinbo Chong, and Nicole Hentrich for workshopping my writing. Thanks as well to The University of Michigan's Asia Library and particularly the help of Liangyu Fu who helped me obtain many of the fascinating government documents and manuals that informed my research.

Thanks to the amazing group of researchers at The Citizen Lab at University of Toronto for helping me with my work and inspiring me with all of their amazing research. Thanks especially to Masashi Crete-Nishihata, Ron Diebert, Jeffrey Knockel, and Lotus Ruan for reading and commenting on early drafts of chapters.

TABLE OF CONTENTS

LIST OF FIGURES

Figure

LIST OF TABLES

Table

ABSTRACT

A large body of literature devoted to analyzing information control in China concludes that we find imperfect censorship because the state has adopted a *minimalist* strategy for information control. In other words, the state is deliberately selective about the content that it censors. While some claim that the government limits its attention to the most categorically harmful content—content that may lead to mobilization—others suggest that the state limits the scope of censorship to allow space for criticism which enables the state to gather information about popular grievances or badly performing local cadres.

In contrast, I argue that imperfect censorship in China results from a precise and covert implementation of the government's *maximalist* strategy for information control. The state is intolerant of government criticisms, discussions of collective action, non-official coverage of crime, and a host of other types of information that may challenge state authority and legitimacy. This strategy produces imperfect censorship because the state prefers to implement it covertly, and thus, delegates to private companies, targets repression, and engages in astroturfing to reduce the visibility and disruptiveness of information control tactics. This both insulates the state from popular backlash and increases the effectiveness of its informational interventions.

I test the hypotheses generated from this theory by analyzing a custom dataset of censorship logs from a popular social media company, Sina Weibo. These logs measure the government's intent about what content should and should not be censored. A systematic analysis of content targeted for censorship demonstrates the broadness of the government's censorship agenda. These data also show that delega-

tion to private companies softens and refines the state's informational interventions so that the government's broad agenda is maximally implemented while minimizing popular backlash that would otherwise threaten the effectiveness of its informational interventions.

CHAPTER I

Introduction: Delegation and Covert Censorship in China

In March of 2018, popular social media company Sina Weibo was thrown into chaos due to a bungled attempt to censor LGBT content on their platform. The controversy began when Sina Weibo announced its intention to ban "gay-themed cartoons, images, and video, citing "the Cyber Security Law of the People's Republic of China." Surprisingly, other social media platforms did not make any such announcements. The public responded forcefully. In China, discussions of LGBT rights are becoming more commonplace and LGBT individuals are becoming more visible. At the same time, the ruling Chinese Communist Party (CCP) rarely addresses the issue. When it does, it is often vague or contradicted by another voice within the Party or government. On Sina Weibo and other social networking sites, users posted hashtags such as #IAmGay and #IHaveGayFriends in protest of Sina's decision. After a swelling of public outrage, Sina Weibo reversed the policy. This reversal was followed by an article in the state newspaper People's Daily that criticized the platform for its decision, asserting that homosexuality was not abnormal and urging regulators to "exercise caution when cleaning up [the Internet] and make sure that they do not confuse [non-illegal content with illegal content] when rushing to take action."[1]

[1]Translation of the editorial can be found here: http://www.webcitation.org/71h6PuJyw

What went on behind the scenes at Sina and in the government that resulted in this sequence of events is difficult to discern. Did Sina Weibo decide to censor LGBT issues in response to a government directive or did they self-censor? Why was this decision not consistent across other social media platforms? Can the apparent reversal of policy be explained by disagreements within government about how to approach LGBT content? Did popular outrage affect the reversal of this decision? While previous studies of censorship have focused solely on the end result: whether content was censored or not, this event suggests that much more is going on in China's system of censorship.

In this dissertation, I break with prevailing explanations in the literature that suggest the regime selectively censors a particular category of content. Instead, I argue that the Chinese Government has broad and expansive censorship goals. Rather than pursuing a selective, *minimalist* agenda for censorship, the regime seeks to covertly implement a *maximalist* agenda for information controls. This maximalist agenda is implemented covertly to reduce invasiveness and visibility of information control interventions. Covert implementation of censorship increases the effectiveness of state interventions and reduces state exposure to popular backlash. Delegation of censorship to private actors facilitates compromise between the public demand for information and the state's preferences. Delegation also shifts blame over censorship to private actors, insulating the state from backlash resulting from unpopular decisions such as Sina's decision to ban LGBT content.

To test hypotheses generated by this theory, I analyze the entire process of censorship, focusing on the many conflicts between individuals in state, society, and the private sector that result in censorship decisions. This analysis shifts the spotlight onto an often-neglected actor in China's system of information control: the private online media platform. State partnerships and delegation to private actors can make censorship interventions more covert and more effective. Private companies are better

equipped to avoid controversy and target their censorship efforts, replacing the state's hammer with a scalpel.

1.1 The Puzzle of Imperfect Information Control In China

Many early works of the impact of the internet on the durability and resilience of authoritarian regimes optimistically suggested that the internet would be a democratizing force. Scholars, politicians, and journalists focused on the liberalizing effect of the internet, claiming that the internet was fundamentally unregulatable. With hardware, software, law, and repression, the Chinese government has tamed much of the internet in China and made early utopian and libertarian visions of the internet seem misguided. Conversely, China's success in the realm of internet information manipulation makes works by Lawrence Lessig seem prophetic. He claimed that eventually, the internet would be bounded and constrained by governments, private interests, and more fundamentally by code—what Lessig considers a form of law in cyberspace (*Lessig*, 1999). In China today, information control regulations are implemented through code which governs and constrains human behavior and "guides the opinions" of China's 773 million netizens.

Certain websites—Facebook, Twitter, Instagram, and Google—are inaccessible behind China's Great Firewall, China's first line of defense against "harmful content" online. Within China's borders, users experience an almost wholly domestic internet that is bounded and controlled. Despite this, the internet within the Firewall is not stagnant, nor is it devoid of a vibrant public sphere. One can find numerous tribes of netizens who have staked claims to their own corners of the web. These tribes invent memes, slang, and form coherent group identities. They can be critical, supportive, or indifferent to the ruling CCP. Some tribes obsess over reality TV, but others organize around local interests, confronting authorities with grievances.

Some launch citizen investigations to expose local government corruption.[2] Others harass celebrities hoping for hush payments, shill for private companies, or do a little of both.[3] Some "voluntarily" participate in patriotic campaigns to spread 'positive energy' in support of the government's agenda[4] while others do so as a means to a government paycheck.[5]

While information controls in China are encoded into law, and physically and virtually embedded into China's network infrastructure, content explicitly deemed harmful by the Communist Party routinely seems to fall through the cracks. In China, a Leninist single-party autocracy, there is sufficient space for non-official organizations to organize and for "harmful content" to spread. A vibrant, though constrained public sphere appears to have emerged in China. This is puzzling because China arguably has the most technologically advanced system of information control in the world. How and why[6] did this happen?

Scholars of the internet in China usually address this puzzle of "imperfect censorship" in one of two ways. The first draws on social movements literature, claiming that imperfect control of the internet results from the "cat and mouse game" between state and citizen over who controls a contested public sphere. The second sees imperfect control as a minimalist government strategy, whereby the state is deliberately

[2]Ai Weiwei famously used the internet to aid in his citizen investigation of corrupt practices that resulted in the deaths of schoolchildren during the Sichuan earthquake of 2003.

[3]China's paid information manipulation efforts are massive, with a large amount of social content created by what netizens call the "Water Army," netizens who are paid to post comments for businesses or celebrities. Some of these individuals will organize coordinated attacks on celebrities hoping to receive hush money, others will post positive product reviews for companies, and some will do a little of both. See http://www.webcitation.org/72Gl4uuBi

[4]Several groups of netizens appear to spontaneously brigade in opposition to China's critics. These netizens are sometimes called the "Volunteer Fifty Cent Party" or the "Little Pinks." During the Taiwanese election, several of these individuals organized a campaign on Baidu Tieba to circumvent censorship of Facebook and post pro-China messages on Tsai Ying-wen, Taiwan's now president's Facebook page.

[5]See Chapter 5 for more on paid regime commentators, also known as astroturfers of the "Fifty Cent Party."

[6]Yawen Lei theorizes that the vibrancy of China's public sphere is an inadvertent consequence of efforts to modernize legal and media institutions within China's authoritarian system. She claims that as a consequence, citizens now have unprecedented opportunities to challenge the regime, organize around law, and influence policy (*Lei*, 2017).

selective about the content that it censors. While some claim that the government limits its attention to the most categorically harmful content—content that may lead to mobilization—others suggest that the state limits the scope of censorship to allow space for criticism which enables the state to gather information about popular grievances or badly performing local cadres.

The "cat and mouse game" argument posits that China's public sphere has emerged and expanded as a result of conflict between state (cat) and citizen (mouse) over who controls a contested public sphere. Early scholars of information control in China focused on the "mouse," claiming that the "democratic" structure of the internet and its technological affordances has empowered citizens to evade the state's information controls and expand non-official public spheres. Traditional one-to-many content dissemination was upended by the many-to-many relationships made possible by the internet, "democratizing" content production and dissemination. New technologies and modes of communication, these scholars argued, reduced the state's ability to set the agenda and shape political preferences (*Diamond*, 2010; *Esarey and Xiao*, 2011).

While these works are optimistic about the relative power of citizens, others point out that the state has more control over the institutions and structures that govern the internet—a unique advantage in the cat and mouse game between state and netizen (*Lessig*, 1999; *MacKinnon*, 2009; *Morozov*, 2012). (*Han*, 2018) claims that in addition to state control over the structures and institutions of the internet, both regime supporters and regime opponents can take advantages of the public sphere that exists in China. Regime supporters can also use online tools and platforms to advocate on behalf of the state. The state's benefits from the support of these social groups as it can artificially increase the visibility of these groups, making it seem as if the official position has broad popular support among ordinary people.

The second explanation for the emergence of China's public sphere argues that censorship is imperfect on purpose, and is a deliberate choice made by the state.

Instead of focusing on state-society conflict, these scholars explore the determinants of what is permitted and what is censored, theorizing about the government logic behind these boundaries. Some claim that the only thing that is categorically off limits is collective action content (*King et al.*, 2013, 2014). Others claim that censorship is strategically selective to facilitate "public opinion supervision" of local government officials, that is, identifying corrupt or poorly performing local cadres through mass surveillance of social media (*Lorentzen*, 2014; *Dimitrov*, 2017). Roberts claims that the state knows it can't perfectly censor, so it relies on "friction" and "flooding" to control access to harmful content. She demonstrates that even though government censorship does not expunge all non-official information, it makes finding this information more difficult. She finds that these efforts are highly effective in influencing Chinese netizens' exposure to information.

1.2 Maximalist Censorship Strategy and Covert Implementation

In contrast to the aforementioned explanations, I argue that imperfect censorship in China results from a precise and covert implementation of the government's maximalist strategy for information control. Despite the apparent imperfections in censorship implementation, the state is intolerant of government criticisms, discussions of collective action, non-official coverage of crime, and a host of other types of information that may challenge state authority and legitimacy. In order for censorship to be covert and more effective, the state adopts methods of implementation of censorship that appear imperfect, but in reality are highly effective because they are precisely targeted, constrained, and hidden. The state covertly advances its maximalist censorship objectives in 3 ways.

First, the state delegates censorship to internet content providers (ICPs). ICPs

serve as a mediator between the interests of the state and the interests of society. Because ICPs weigh the costs of defying government directives with the benefits of satisfying market demand for information, censorship outcomes reflect a compromise between the state's maximalist censorship objectives and society's demand for information. ICPs strategically respond to market demand for information, the actions of competitors, estimations of the state's capacity to monitor compliance, and the expected sanctions for non-compliance. This results in a maximization of the breadth of censorship under market-informed constraints of what level of censorship society will tolerate.

Second, the state targets information control and repression selectively to influential social actors. Selectively targeting information control, repression, and cooptation to influential users reduces the average citizen's propensity for experiencing information control. At the same time, because influential individuals generate most of the content online, this strategy achieves a sizable reduction in the visibility of harmful content. This strategy works because it is incentive-compatible with ICPs' desire to reduce disruptiveness of information controls to their users. The government can rely on social media companies to identify and report influential users to them and to prevent censorship of users from being widely observed.

Third, the state engages in widespread astroturfing—production of pro-government content by government employees who are masquerading as "grassroots" individuals—to covertly influence perceptions of popular support for the state's positions. Astroturfing suppresses the speech of individuals with non-official opinions and artificially increases the share of official opinions. This is accomplished covertly—without implicating the state as the propagandist behind the content. Even though some of this content may be identifiable, the state can plausibly deny authorship.

While it may appear that information controls are incomplete when observing the outcomes of interventions, they are incomplete for a reason. The state benefits

from the selectivity of delegated censorship. Delegation to private companies hides censorship, digital repression, and opinion guidance from public view. Delegation also allows social media companies to serve as a mediator between state and society that prevents direct state-netizen conflict. Private companies such as Sina Weibo advocate for user demand when possible, often deliberating and negotiating over censorship directives with leadership. This is because providing users with content they crave and preventing an overly-censored environment is good for their bottom-line. Because most citizens are not involved in these conflicts, and do not observe much of the state's censorship, the internet feels relatively free and unconstrained to most netizens. This benefits the CCP in three ways:

First, hidden censorship allows the state to engage in "public opinion supervision," the process of mining social media data to identify and respond to popular grievances, without too much fear of preference falsification.[7] Alerting users to the censorship and information controls that happen routinely on their platform could poison the well from which the regime draws insights about public opinion. In this way, delegation of information control to private companies can mitigate information problems inherent in authoritarian rule (*Wintrobe et al.*, 1998; *Wallace*, 2014; *Dickson*, 2016). Second, hiding information control from users not only benefits social media companies by reducing the cost of censorship implementation and improving user experience on their platforms, but it also benefits the state by reducing the likelihood of direct state-netizen conflict. If a netizen who is a regime supporter or who has neutral beliefs about politics finds out she has been censored, she might reassess her support for the regime. If a regime-opponent finds out she has been censored, she may escalate her anti-regime behavior, or attempt to circumvent censorship. Third, outsourcing censorship to private companies gives the government a scapegoat when censorship decisions result in popular backlash. If the public displays moral outrage at a censorship decision,

[7]The theory of preference falsification suggests that individuals will publicly state preferences they find suboptimal, because their optimal preference is less socially acceptable (*Kuran*, 1987).

the state can order a change in policy and claim that the social media company misinterpreted directives.

1.3 A Brief History of Information Control In China

1.3.1 Pre-Internet Information Controls

In the Chinese political system, the media is often described as the "mouthpiece of the Party and the bosom-friend of the people."[8] Borrowing from Marxist-Leninist theory on media, the CCP has considered control of mass media a pillar of its rule. Throughout its history, however, the methods of Party control over mass media, the structure of mass media, and the technologies of information dissemination have changed.

During the Chinese Civil War, the CCP exercised strict control over information within the People's Liberation Army (PLA) predominantly by means of violence. This was reflected in Mao Zedong's concept of the "mass line," which described a process whereby leaders were responsible for gathering the chaotic ideas from the masses, systematizing them, and retransmitting this corrected version through propaganda and "thought reform." The "correct" way of thinking was policed by leaders through physical violence and fear-fueled indoctrination. At communist bases such as Yan'an, uttering a wrong word might result in public criticisms or in some cases displays of physical violence. Party members were encouraged to confess their impure thoughts and actions in mandatory "self-criticisms"–written admissions of errors in their thoughts or behaviors. At the same time, they were expected to police each other's thoughts and inform on others. This climate of fear intensified during the "Rectification Campaign" which was in party carried out in response to articles in the

[8]This phrase is often used by state media during anniversaries or when proclaiming their allegiance to the party. One might also hear this phrase uttered in mandatory Marxism classes at Chinese universities. Chinese: 媒体是党的喉舌，人民的知音

official paper of the PLA that were critical of a "system of hierarchy and privilege" in the Yan'an base (*Chang and Halliday*, 2005, 266-269). This climate of fear regimented the minds of the early Communists and ensured that leaders alone—especially Mao—were empowered to communicate with the masses through major Party newspapers, the Liberation Daily and the People's Daily. These writings served the purpose of mobilizing the masses and lower level cadres through published speeches and writings of leadership.[9]

Within a few years after Mao Zedong proclaimed the People's Republic of China, all media organizations were either shuttered or subsumed by the CCP, resulting in a Party monopoly over media. Until media marketization reforms, traditional media were almost entirely state-owned. Editors and staff of media organizations were also party officials. With no competition and no alternatives to state media, there was very little pressure to respond to audience demands. This meant that the Party could exercise control over media organizations through Party membership and the nomenklatura system (political appointments, promotions and demotions). To further pressure editors and journalists, the regime at times used the threat of violence to enforce compliance and promote self-censorship.

Beginning in 1992 and accelerating under Hu Jintao, media in China began to commercialize. This change was a response to both general trends of increased personal choice and freedom in Chinese society, and a realization on the part of leadership that continued state subsidies to media organizations were unsustainable. After media reforms, the number of publishers in China increased, and for the first time, some of them were outside of the Party's direct administrative control. Because some publishers were commercialized or semi-commercialized, the state could no longer rely on political appointments to control publishers, and were confronted with a problem: what is politically sensitive is often a big seller at the newsstand. To control

[9]These speeches served the purpose of mobilizing the masses in support of Mao's Rectification Campaign. For more on information controls in Yan'an, see *Volland* (2003)

commercialized media, the CCP began to issue publishing licenses, which could—in extreme circumstances—be revoked if publishers failed to report within the acceptable bounds.[10]

Though the Party loosened its control over media, the number of publishers remained tractable and could be controlled through publishing licenses and legal/violent threats. The number of licenses could be selectively restricted by the Party and could include stipulations on what kinds of reporting were allowed. Furthermore, although the Party could not directly fire editors or journalists, they could weaponize the legal system against editors and journalists who cross the line. For example, in 2004, Nanfang Media Group published reports of the abuse and death of migrant college student Sun Zhigang at the hands of state security. This story resulted in public displays of moral outrage online and in the streets. Guangzhou provincial and municipal authorities retaliated against the commercial media conglomerate by launching an investigation into their finances, arresting 3 top executives, and sentencing the editor-in-chief to 12 years in prison. These highly visible displays of state repression become seared into the minds of reporters and editors, warning those in commercial media of the consequences of crossing the line.[11] Though occasionally commercial papers do push the bounds of acceptable reporting, Daniela Stockmann has found that overall, the reporting in commercial and state-owned media is synchronized, an indication that the pre-reform control structures have more or less adapted to commercialization of media in China (*Stockmann*, 2013).

[10] "Edge-ball" strategies are common in the Chinese media industry and are used to increase readership. Edge-ball comes from ping-pong, where shots that target the edge are most likely to be winners, but also a much more risky move for the offensive player (*Keane*, 2001). The further a corporate agent is from a government actor's locus of power, the more they tend to behave in ways that are on the edge of what a government actor deems appropriate (*Stockmann*, 2013).

[11] *Stern and Hassid* (2012) call these displays of state repression "control parables." These control parables lead media practitioners to self-censor out of fear that extreme consequences might follow from pushing boundaries or failing to cooperate.

1.3.2 Post-Internet Information Controls

Control over media became much more intractable with the rise of the internet as a major source of information. In its efforts to control traditional media, the Party had relied heavily on pre-publication audits of information through communication between editorial staff and propaganda departments (*Brady*, 2009). The ability to control publishers became difficult in the new information regime brought about by the internet (*Esarey and Xiao*, 2011). Whereas before information was transmitted from publisher to consumer in a one-to-many relationship, the internet blurred the lines between publisher and consumer facilitating many-to-many information sharing relationships. The state could no longer exercise the same level of control over publishers because suddenly, everyone with an internet connection could become a publisher. Instead, the Party shifted its focus toward post-hoc censorship and delegating censorship of content to online platforms.

The process of delegation of censorship to ICPs in China began with two guidelines issued by the State Council.[12] The regulations stipulate that "business websites" are liable for the content on their own sites, and are required to police and remove illegal content. It also specifies penalties for noncompliance, stating that, "If the case is serious, it shall order the perpetrator to suspend operations and undergo rectification or to temporarily shut down its website."[13] This regulation mirrors the system that was already in place for traditional media, where editors were liable for the legality of content they published. Because of the structure of the internet, however, propaganda departments and other existing bureaucracies could not rely on pre-publication censorship as it had in its control of traditional media.

The early years of the internet in China involved the state's shaping of the ICP

[12] "The Means of Managing Internet Information Services" (互联网信息服务管理办法), archived here http://www.webcitation.org/72GlTQoOO and "The People's Republic of China Telecommunication Regulation" (中华人民共和国电信条例) archived here: http://www.webcitation.org/72GlUunpl

[13] Official English translation archived here: http://www.webcitation.org/72GlWKsm7

market through these regulations. From 2000 to 2009, the internet in China was social, but fragmented into smaller platforms such as blog service providers (BSPs) and bulletin board systems (BBS). Each platform hosted only a handful of bloggers with large followings. Online platforms that shared "harmful content" were warned by authorities and put on notice by the "China Internet Illegal Information Reporting Center," their internet licenses presumably in jeopardy. With pressure to clean up their platforms, many built or purchased software platforms to censor their own content. Those who failed to do so were shut down. The popular BSP 'Bullog' famously met this fate in 2009 for hosting "harmful comments on current affairs."[14]

This all changed when "micro-blogs," and more specifically, Sina Weibo, gained sudden and massive popularity in late 2009. Sina Weibo united netizens once siloed in small BBS or BSP platforms in a large network that enabled a "loose-ties" type of content sharing.[15] While producing content for mass consumption on BSPs required writing skills and a large following, Sina Weibo facilitated a low-cost way for ordinary people to interact with each other, dramatically increasing the number of content producers on the Chinese internet (*Cairns*, 2016b). With massive popularity, and a competitive share of the social media market, the government risked a public backlash and loss of domestic market share if they shut Sina Weibo down or controlled the platform too strictly (*Pan*, 2016).

Traditional media regulations and institutions failed to adapt quickly to this wholly new mode of information propagation introduced by Sina Weibo and other microblogs. While smaller platforms could be pressured to self-censor and could be shut down under "serious" circumstances, social networking sites like Sina could not reasonably be expected to perfectly police all content produced by its hundreds of

[14]News report on the shutdown archived here: http://www.webcitation.org/72GlzujAz

[15]Loose-ties networks involve the sharing of content through loose social connections, i.e. people you may not personally know such as a journalist, celebrity, or politician. Loose-ties networks can usually also accommodate close social connections such as family, friends, or coworkers. For an overview of these different types of social networks in China, see (*Stockmann and Luo*, 2017).

millions of registered users. Even today's most cutting-edge automated natural language technologies for identifying objectionable content have serious limitations. At the time of Sina's rise, there were no market-ready tools to aid Sina in its censorship responsibilities. Every individual internet user suddenly became a publisher, but did not have to run what they published by the CCP censors before pressing send. Without pre-publication control over "harmful content," these new social media sites quickly became a space outside of the traditional system of information controls where citizens could push the bounds of allowed public expression.

The Chinese government faced a dilemma. The internet promised to bring growth and productivity to China, but the internet's architecture and these new platforms made it nearly impossible to use traditional information control tactics. Adapting existing bureaucratic structures to such a fundamental change would have been difficult and costly. New modes of control were necessary, but the bureaucratic structures needed to control the negative externalities of this new "information regime" had never existed before.

Instead of building these structures from the ground up, the problem was addressed experimentally[16] with the central government directing the process[17] through broad mandates such as "prevent public opinion emergencies" and "guide public opinion." Bureaucracies and governments responded improvisationally to accomplish these broad tasks. Eventually, these specialized modes of information control were

[16]In the Chinese system, experimental policymaking involves decentralization and autonomy in early stages of policy development, and then as policies are tried and tested locally, they are adopted by the central government and rolled out nationally in a process called moving "from point to surface" (由点到面) (*Heilmann*, 2008)

[17]The process of experimentation that characterized the development of internet regulations resembles a process Yuen Yuen Ang calls "directed improvisation." Directed improvisation is an experimental and decentralized approach to policy making in the absence of any guiding precedent or strong institutions necessary to achieve relevant policy goals. In the realm of foreign direct investment and early economic reforms, (*Ang*, 2016) identifies an campaign-driven, evolutionary process of institution building, which she refers to as "beehive campaigns" of development. In early stages of development, each agency (bees) were enlisted to "prospect for investors for their home states" (gathering honey) while simultaneously performing "formal functions (e.g., environmental protection, law enforcement, personnel management). Similar campaigns exist in China's system of information control, particularly with regard to identifying and responding to public opinion emergencies.

adopted by the State Council and communicated as national templates (*PRC State Council General Office*, 2016).

1.4 How Corporate-Delegated Censorship Works

Today, censorship is delegated to corporations through directives that vary in their levels of specificity: from vague and broad to precise and targeted. According to a series of interviews conducted by Christopher Cairns, a wide range of regulatory bureaucracies, government organs, and individual leaders at all levels of government participate in this process. A former employee in the censorship division of Sina Weibo—and the source of the leaked censorship data used in this dissertation—describes the process in detail. A "government relations specialist" communicates with these myriad government actors receiving "clear, direct and urgent" orders about "whose accounts need to be removed and which posts need to be deleted." This individual also "frequently goes to government meetings" to understand "broad censorship guidelines," which result in "vague" censorship orders (*Wang*, 2016b,a).

Cairns (2016a) found that the responsibility of censorship is fragmented across many bureaucracies, making "life more complicated for Internet companies in deciding whose orders to follow." As one company insider claimed, the system was "a mess." The leaked censorship documents used in this dissertation confirm what he found, through interviews with insiders, to be the basic bureaucratic structures that monitor and direct censorship at Sina Weibo.

Today, social media companies in China are managed at the local level. This means that, for the most part, delegation, monitoring and sanctioning of private companies are carried out locally. Because Sina Corp. is headquartered in Beijing, regulatory authority over Sina Weibo is mainly concentrated in two Beijing provincial-level bureaucracies. They are referred to as the Beijing Municipal Internet

Propaganda Management Office (Internet Management Office)[18], and the Public Information and Internet Safety Supervision Department of the Beijing Public Security Bureau (Supervision Department).[19] Alongside these two main bureaucracies Sina takes direct orders from the State Council Information Office,[20] which is a national-level bureaucracy that directly reports to the State Council, China's top government agency. Several other bureaucracies make direct requests to Sina Weibo, including provincial propaganda bureaucracies such as the Shanghai News Information Office, local internet police at all levels of government, provincial governments such as the Tibet Autonomous Region, and area-specific bureaucracies such as The State Administration for Industry and Commerce. Aside from bureaucracies, directives also come indirectly by way of six Sina Corp. managers who appear to respond directly to lobbying from unnamed individuals. Note that the data in this dissertation predates the rise of the Cyberspace Administration of China (CAC) as the chief regulator of China. In Chapter 3, I discuss some changes to how the system works after the CAC's rise to prominence.

As I will stress in later chapters, though many bureaucracies have the power to send directives to Sina Weibo, Sina Weibo is ultimately responsible for pressing "delete." In many cases, for a variety of reasons, Sina does not do so.

1.5 Data

The empirical work in this dissertation draws upon two sources of data. The first is a dataset of leaked company censorship logs from Sina Weibo which I manually compiled and coded with several research assistants. The second is a combination

[18]In Chinese: 北京市互联网宣传管理办公室, abbreviated as 网管办 in the logs.

[19]In Chinese: 北京市公安局公共信息网络安全监察处, abbreviated as 网监 in the logs.

[20]The SCIO was elevated in 2011 to directly report to the State Council and was renamed the SIIO. The logs still use the name SCIO, so to avoid confusion I will use SCIO, even if technically the organ was called SIIO at the time. In Chinese: 国务院新闻办公室, abbreviated as 国新办 in the logs.

of two datasets that measure the outcome of censorship on the Sina Weibo platform during the period of time the leaked censorship logs were made. This allows us to examine the deliberation and contestation between government and Sina Weibo leading up to censorship decisions as well as a measure of the end-result of these deliberations (the censorship outcomes).

1.5.1 Leaked Censorship Logs

In early 2016, the Committee to Protect Journalists reported on a leaked cache of documents from Sina Weibo's censorship office.[21] These documents log government directives, company policies related to content moderation, and management decisions about how to proceed with censorship of content. The logs record this information so that it can be shared with employees working in different shifts, minimizing duplicated management effort. This dataset is the first of its kind to provide a look into how censorship delegation works in practice, describe contention between governments and private social media companies in China, and show both what is censored and what the government intent behind censorship was. What emerges from the logs is not a picture of tight control over information, but one in which a fragmented and decentralized government struggles to compel companies to enforce broad informational goals.

These data, in their raw form, consist of 588 Microsoft Word documents, each containing dozens of individual logs. Logs are notes related to government directives, management censorship decisions, work guidelines, employee duties, etc. In total there are 8,427 unique logs. Most logs include a mention of a certain type of content and instructions on how to proceed with censorship, disseminating management decisions about censorship to employees.

[21] I am not in direct contact with the source, but the source has been vetted by journalists at CPJ and I have communicated with the source through contacts at CPJ. The source has consented to the publication and use of these data for researchers.

There are 3 main varieties of internet censorship in China. The first, domain-level filtering, blocks access to certain websites such as Facebook or Twitter. The second, keyword-based filtering, automatically blocks, or automatically triggers surveillance or review of content if it contains one, or a combination of blocked words. Third, manual review, often triggered by keywords, is the process of sending posts to humans employed by an internet platform to make manual decisions about whether or how a post should be censored. While many datasets related to censorship capture only keyword-based censorship[22] or manual content review,[23] log data include logs about both. This dataset includes logs from 2011 to 2014, overlapping with other studies of censorship in China (*King et al.*, 2013, 2014).

There are limitations to the external validity of inferences drawn from these data, as these logs are from a single social media company, Sina Weibo. That being said, certain logs suggest that these inferences can generalize to other social media companies, particularly Tencent. Several logs indicate that competitor Tencent receives the same directives from the same Beijing provincial-level bureaucracies as well as the State Council Information Office, and more recently the Cyberspace Administration of China (CAC). Tencent, like Sina, also delays implementation or disobeys directives when doing so gives them a competitive edge. Additionally, the market capitalization of both Sina and competitor Tencent were similar at the time these logs were written, which means both had similar leverage when it came to negotiating government directives. It is reasonable, based on the data from the logs to expect that the process of reporting of users to the authorities is similar at Tencent. Together, Sina and Tencent represent the vast majority of the social media market. The logs also indicate that tech companies in China cooperate or are influenced by each other's censorship

[22]Keyword-based censorship is usually a website's first line of defense. Keywords are used to automatically block certain searches that include sensitive terms, automatically trigger surveillance if a user uses certain terms, or automatically stage content for human review. Oftentimes these lists will leak and are then analyzed by researchers. See (*Crete-Nishihata et al.*, 2017; *Ruan et al.*, 2016; *Ng*, 2016; *Knockel et al.*, 2017, 2011, 2015)

[23]Manual content review is explored in an experimental setting in *King et al.* (2014).

decisions. For example, Sina Weibo cooperates with Baidu when developing keywords for blocking and content filtering.[24]

Even a conservative approach to inferences drawn from these logs does not diminish their significance. Sina Weibo is a very large and popular social media company in China. During the time of the logs it was the most popular microblog and was ranked in the top 3 of all domestic social media companies by monthly active user statistics. During the time of the analysis, Sina Weibo boasted over half a billion registered users.

This unique dataset, which captures the population of logs from 2011-2014, allows for direct measurement of the intention and goals governments and private internet companies. In particular, these capture the decisions of Sina, parent company to the Sina Weibo platform, and Tencent, a Sina's chief competitor. Logs include 1) data on the source of a censorship request, and 2) Sina's (and sometimes competitor Tencent's) response to this request. This allows us to disentangle the intentions of government principals and corporate agents.

The data contained in these logs confirm earlier work on the fragmentation inherent in the Chinese political system (*Oksenberg and Lieberthal*, 1988) and more specifically the Chinese media and propaganda system (*Cairns*, 2017; *Stockmann*, 2013; *Ang*, 2014). Logs explicitly name several government principals at different ranks and bureaucratic functions. They also reveal that multiple corporate agents compete with each other, inform on each other, and calibrate compliance based on observed compliance of competitors.

[24]In a log from May 7, 2013, Sina instructs employees to add to a shared database of blocked keywords with Baidu in accordance with an arrangement with the company. The log reads: "In the future, if you add level A, B, or C blocked search keywords, please also add these keywords to the Baidu cooperation data control system. The specific correspondence is, level A blocked keywords correspond to Baidu's level A, and B and C level blocked keywords correspond to Baidu's level B."

1.5.2 Measures of Censorship (Free Weibo and Weiboscope)

As I measure of censorship outcomes, I process data from GreatFire.org and Weiboscope, two projects that measure censorship on the Sina Weibo platform. The GreatFire.org dataset includes 47 million weibo posts spanning from 2009-2018 and the Weiboscope dataset includes 226 million posts from only 2012. Using these data I can compare the instructions in the logs to actual censorship outcomes at the event level. Both datasets are constructed by recording a post soon after it has been created and later querying for that post at fixed intervals to see if the post has been censored or deleted.

1.6 Roadmap

In this dissertation, I draw upon these data to answer several key questions about information controls in China.

In Chapter 2, I test the hypothesis that outsourcing censorship to private companies results in more "covert" censorship that is designed to hide censorship from the end-user. I argue that covert censorship improves user experience and reduces costs of censorship implementation by preventing users from escalating their behavior or circumventing censorship efforts. Covert censorship improves the state's ability to guide opinion and demobilize opponents because it reduces the occurrence of public conflicts between state and society while accomplishing the goal of limiting the scope and spread of counterhegemonic discourse. Because media platforms wish to minimize the amount of work necessary to carry out censorship directives, they avoid what *Roberts* (2018) calls "backlash," angry or defiant behavior resulting from making netizens aware that their posts have been censored.[25] When users know what is being censored, they often attempt to circumvent censorship by using homophones,[26]

[25] *Roberts* (2018) found that when users find out they have been censored, they often continue posting about the off-limits topic, escalate their behavior, or attempt to circumvent censorship.

[26] Homophones are words in Chinese that sound nearly identical but use different characters. For

homoglyphs,[27] or by reposting from secondary accounts. To add empirical support to this theory, I manually label a large database of leaked censorship logs from Sina Weibo by censorship type (covert vs. overt). I find that the vast majority of censorship on the Sina Weibo platform is covert (nearly 80%), and that it is favored because it reduces the workload of content moderators, and thus reduces costs to private media companies. I also find that covert censorship is used more frequently on more influential users, an indication that social media companies are concerned with popular backlash. Qualitative evidence from the log corpus suggests that covert censorship is a successful method of demobilizing regime opponents.

In Chapter 3, I address the puzzle of imperfect censorship: how China can simultaneously boast the most extensive and advanced system of censorship in the world, while its internet platforms are riddled with criticism, contention, and heated political discussions. Popular consensus in the political science literature suggests that authoritarian governments have a minimalist censorship strategy, and strategically "allow" specific categories of content online. In this chapter, I argue that imperfect censorship is mostly a result of principal-agent problems between delegating government principals and profit-driven corporate agents such as social media companies. While social media companies can be punished for failing to comply with government directives, they can also profit from sensational, anti-government content. Social media companies push back on censorship directives to satisfy market demand, implementing a maximalist censorship strategy that is constrained by the market. I find evidence to support this theory from a corpus of leaked censorship logs from China's

example, when a pro-reform document "Charter 08" or 零八宪章 (pronounced líng bā xiàn zhāng) was blocked, netizens began to use the phrase "County Magistrate 08" or 零八县长 (pronounced líng bā xiàn zhǎng) in its place.

[27]Homoglyphs are characters that look like one another such as "已" and "己" or "因" and "困" or "日" and "曰." To circumvent keyword censorship, netizens often replace the blocked character with another character that looks similar. For example, during anti-Japanese protests, if the phrase "反日游行" or "Anti-Japanese Demonstration" is blocked, users might write "反曰游行" or "Against Saying 'Demonstration"' instead. Note in the second phrase, the second character "曰" has a middle horizontal stroke that is not connected to the right vertical stroke, in contrast to the character "日" in the first phrase.

second most popular social network, Sina Weibo. In censorship logs—internal records of government directives and company implementation instructions—I find that 16% of all directives are deliberately disobeyed. Many logs in these data include Sina's explanations for disobedience, and their motivations are clear—they want to gain an edge over competitors by providing more compelling information in spite of it's level of sensitivity to the government. Even when Sina Weibo chooses to implement government directives, I find that market-forces drive the company to resist censorship. By retreiving exact posts targeted for censorship from hundreds of millions of historic Sina Weibo posts, I calculate the rate of Sina's implementation of instructions to employees. I find that the rate of censorship implementation at Sina Weibo dramatically increases when content directly affects Sina Weibo's share price, suggesting that though Sina has relatively high capacity to censor, they deliberately shirk when implementing government directives, likely in an attempt to satisfy user demand for information.

In Chapter 4, I test a prominent theory of minimalist censorship: the "collective action potential' hypothesis of (*King et al.*, 2013). The collective action potential hypothesis posits that the Chinese government intends to censor collective action but not government criticism. Because the log corpus directly captures intentions of government actors, I manually labeled all 8,427 logs by topic category. The resulting distribution of topics suggests that the government has a maximalist censorship agenda. While criticism, discussions of government leadership, and crime are targeted much more frequently than collective action content, it appears that the scope of China's censorship targets is much broader than we thought. In a separate working paper with Mary Gallagher, we offer support for an alternative hypothesis, that social media governance involves precise targeting of *individuals*, not content. The state targets users to repress or co-opt based on their influence and clout—or in network parlance, centrality of their node in counterhegemonic subnetworks—regardless of the

content they are posting.

In Chapter 5, I outline a method for automatically detecting a type of covert information control called astroturfing. Using various machine learning and information retrieval methods, I find that nearly 15% of all comments in Chinese news media's comment sections are created by "astroturfers," individuals being paid by the state to produce pro-government content while appearing to be ordinary citizens. I find that astroturfers in China work at a wide range of bureaucracies and are part of larger public opinion management teams tasked with "public opinion supervision," "guiding public opinion," and "preventing public opinion emergencies." This method of detecting astroturfing will be used in future work to measure the effects of this information control tactic and analyze the content and strategy behind astroturfer messages.

CHAPTER II

Covert Censorship

2.1 Introduction

Sharing something on social media that no one interacts with can be an anxiety-inducing experience, prompting reflection on whether the content was as insightful or amusing as initially thought. Receiving no likes or retweets, however, doesn't necessarily indicate that content is unpopular. It could also mean it was censored. Using a new dataset of internal censorship documents from a popular social networking site, I show that a majority of Sina Weibo users in China who experience censorship will not know they have been censored. They will instead observe their posts languishing on their timelines without likes, upvotes, or comments.

In this chapter, I will begin by introducing the concept of "covert censorship": censorship that is not visible to the person being censored. I will then outline how the profit incentives of private social media companies lead them to prefer covert rather than overt censorship tactics. I then examine the impact of covert censorship on users using a leaked database of censorship logs from Sina Weibo that record whether a post was overtly or covertly censored. I find that Sina Weibo instructs content moderators to use covert censorship tactics in the vast majority of logs (79.27%). I hypothesize that the purpose of these covert censorship tactics is to prevent the user from finding out that they have been censored for three reasons: to diminish users

capacity to circumvent censorship, to prevent backlash that comes from censorship, and to help platforms appear relatively free and open. Analysis of log data support this hypothesis, but further data from survey experiments is needed to understand how covert censorship tactics affect user behavior.

2.2 Background

In this section I introduce how censorship works in China and at Sina Weibo in particular. I begin by situating the censorship observed at Sina Weibo within the many high-level methods of censorship used to limit free expression in China. I then explain how the process of censorship works at social media companies and introduce the main methods of censorship used by content moderators at Sina Weibo.

2.2.1 Types of Censorship

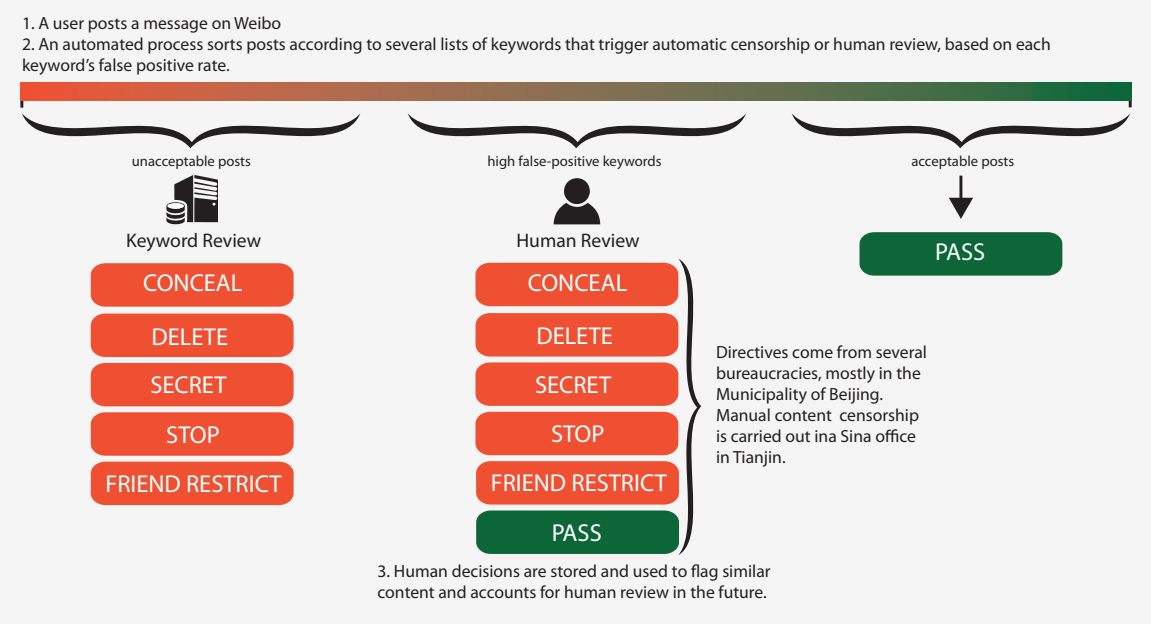
Censorship in China comes in many distinct forms. At a high level, there two distinct varieties of censorship: micro-censorship and macro-censorship. Macro-censorship uses software- and hardware-based interventions that prevent whole domains from being accessed.[1] Macro-censorship is often referred to as "The Great Firewall of China," and is what makes foreign websites like Facebook, Twitter, and the New York Times inaccessible. By contrast, micro-censorship targets individual expression within the subset of websites one can readily access in China. Micro-censorship happens at the post, comment, or article level and is carried out by social media companies or other internet content producers (ICPs).

Micro-censorship is accomplished in several ways: keyword filtering, manual content review, and algorithmic filtering. There are a wide range of micro-censorship methods, due to the practice of delegated censorship that takes place in domesti-

[1]The Chinese government detects and censors content using deep-packet inspection (DPI) (*Wagner*, 2008, 2009). They also make use of DNS poisoning to prevent routing of traffic to the right IP address.

cally licensed internet companies in China. As such, each platform develops its own methods of combating "harmful" content targeted by government directives. Keyword filtering is the first line of defense of many ICPs. Keywords, either single words, phrases, or co-occurring words or phrases, prevent searches, posts, or comments about the most categorically off-limits content (i.e. links to pornography websites, mentions of the banned cult Falun Gong, mentions of the Tiananmen Square Protests, or mocking nicknames of leaders). Keyword lists, however, do not always result in automatic blocking or search restrictions. Instead, there are usually different levels of keyword lists. At Sina Weibo, there are 3 levels of keywords based on some combination of political sensitivity and the likelihood that a non-sensitive post will be flagged by a user who includes that keyword (false positive rate). The list with the most sensitive/lowest false-positive rate keywords will be used to automatically censor content that includes these keywords. The other two lists trigger the second kind of censorship: manual content review. This process is visualized in Figure 2.1.

Figure 2.1: How Sina Censors Content



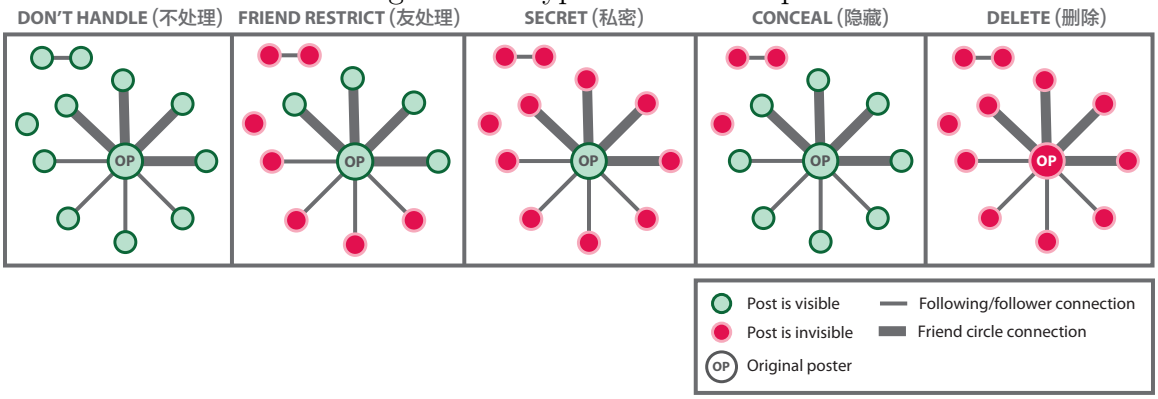
Manual content review is a process where social media companies send content to content moderators who then manually decide whether or how the content should be censored. Decisions about censorship are sometimes saved in a "sample database" (样本库) to detect similar content in the future. This is useful when users try to evade censorship by using homophones, homoglyphs or other censorship circumvention methods because blocking based on the "sample database" is informed by similarity measures rather than simple keyword matches. This is called algorithmic review, which according to the source of the leak was under development toward the end of the log data and did not work very well.[2]

2.2.2 Types of Censorship at Sina Weibo

At Sina Weibo, employees in charge of content moderation make most of the decisions about what is and is not censored on their platform. Posts are flagged and sent to these content moderators through keywords, management-directed "audits" (审核) of certain posts or accounts, or user reports. Though these content moderators do outright delete content, they have a variety of other methods of "handling" (处理) content. Employees working at Sina Weibo commonly choose to "handle" content in one of 5 ways: "delete," "secret," "friend restrict," and "conceal." All of these ways of handling content involve hiding content from a subset of users on the Sina Weibo platform; they are visualized in Figure 2.2 and are described in Table 2.1.

[2]The source said, in an interview with CPJ: "The department had plans to computerize censorship, designing programs to enable computers to complete complex censorship tasks. But the plans didn't pan out, mainly because they were too expensive. The cost of manual labor in comparison was lower. Sina's image recognition technology, to make computers identify the content of pictures, in my opinion was pretty bad, but it was being developed."

Figure 2.2: Types of Censorship



Though the five censorship choices mentioned above are the most commonly used in the data, they are not an exhaustive set of the ways in which Sina handles objectionable content. Oftentimes, rather than getting rid of content, Sina Weibo will choose to slow the spread of that information, increasing what *Roberts* (2018) calls "information friction." The main way in which Sina Weibo does this is by "stopping functionality." This usually entails disabling sharing features such as private messages or retweets/reshares. Sina also at times cooperates with the authorities by sharing information on public opinion or informing on its users' bad behavior. Certain users and posts can be "reported up," or escalated to a "government affairs liaison" and potentially forwarded to the authorities (*Gallagher and Miller*, 2018). Sometimes users will have posts deleted and will be warned in a private message that their post contained objectionable content. Users can also be banned for any fixed period, or have their account deleted. Oftentimes users will create new accounts after they are banned and continue posting objectionable content. These users are referred to by Sina as the "reincarnation party" (转世党). These individuals, when identified, are sometimes IP blocked.

2.3 Covert Censorship

I define covert censorship as measures to limit the visibility of content in a way that is intended to be unobserved by a particular individual or group (usually an author or searcher). One common type of covert censorship measure is called "shadow banning," a type of censorship that makes content invisible to all but the original poster. Shadow banning is used in many social media companies inside and outside of China.[3]

Covert censorship has become more and more common in China in the past several years. On China's most popular social media platform, WeChat, shadow-banning is the default method of censoring content in group chats and one-on-one messages. The Citizen Lab at University of Toronto has also found that shadow-banning affects Chinese users and foreign users differently.[4] In late 2016, Baidu, China's equivalent of Google, stopped notifying users that their search results "may relate to content that does not comply with relevant laws, regulations, and policies, and have not been displayed."[5] Sina Weibo stopped including a similar notice in its search results in 2014.[6]

For content posted on a user's timeline at Sina Weibo, there are three varieties of covert censorship: secret, friend restrict, and conceal. "Secret" is equivalent to shadow-banning as it is commonly understood; it hides content from all users but the original poster. "Conceal" and "friend restrict" restrict the visibility of content to different levels of social connections but keep content visible to the original poster. "Conceal" limits visibility to the original poster and their friends and followers, "friend

[3]Twitter, for example, uses shadowbanning to combat harassment and "bad-faith actors" though they insist that, by their own definition, they do not shadow ban. Shadow banning is defined by Twitter as "deliberately making someone's content undiscoverable to everyone except the person who posted it, unbeknownst to the original poster." Conveniently, because of one word—undiscoverable—Twitter can claim to not shadow ban according to this definition. Instead, Twitter penalizes "bad-faith actors" by ranking them lower so that they appear at the bottom of a follower's list of tweets, unbeknownst to the bad-faith actor.

[4]See report here: http://www.webcitation.org/71xsYmDvk

[5]A blog documented the change here: http://www.webcitation.org/71wjOx1Aj. Additionally, the change is documented on question site Zhihu here: http://www.webcitation.org/71wjTfffo

[6]See report here: http://www.webcitation.org/71xoUQW0Q

restrict" limits visibility to the original poster's "friend circle" (好友圈). Because most follower/following relationships are weak (the user does not know followers/fans personally, i.e. celebrities, writers, journalists, etc.), Sina Weibo allows users to create a "friend circle" that includes individuals with whom they share strong ties such as friends and family. These three types of covert censorship are visualized in Figure 2.1.

2.4 The Market Logic of Covert Censorship

Why does Sina Weibo covertly censor content? Why not just delete all offending posts outright? Some theories of censorship suggest that the topic of content targeted for censorship is important in determining outcomes. Some posit that censorship is selective for informational reasons, informing the center of local government corruption and malfeasance (*Egorov et al.*, 2009a; *Lorentzen*, 2014; *Malesky and Schuler*, 2011). Other works claim that the government deliberately targets collective action content and tolerates government criticism. I hypothesize that content does not influence whether a post is overtly or covertly censored. Instead, I argue that Sina Weibo wants to prevent users, especially highly influential users, from discovering they have been censored. This is due to three main ways that covert censorship prevents alerting users to censorship on their platform and consequently harming their business interests:

First, covert censorship diminishes users' capacity to circumvent censorship because users are not notified of censorship. When users are notified of censorship, they can learn the decision rules governing what is censored, and can exploit loopholes in these rules. This makes the process of censorship adversarial, and Sina Weibo must spend more time and money to moderate content produced by these users. Users can circumvent censorship by changing a few keywords, deliberately misspelling censored words, or using clever phrases that seem innocuous, but carry hidden meaning be-

neath them. This is how users broke the news of Zhou Yongkang, former Politburo Standing Committee member's demise. Because his name was blocked, users referred to him with the name of a popular brand of instant noodles, "Master Kang," that shared a character in his name. When he was expelled from the Party and government, netizes wrote, "Master Kang has been cooked." This put the ICPs in the awkward position of policing the discussion of instant noodles for a time.[7] Censorship circumvention like this can put Sina at risk of government sanctions for violating directives. Keeping track of the many clever wordplays to evade censorship is a difficult task. According to the source of the leak, Sina Weibo struggled with the proliferation of keywords such as these that attempted to get around censorship:

> "During my time at Sina, sensitive words increased from 2,000 to at least over 10,000. Phrases like "McDonald's" and "combo No.3" [code words for organizing protests] became sensitive words during the Jasmine incident, but they didn't get taken off from the "sensitive words list" until the end of 2012, a result of both tardiness and playing it safe. The ever-expanding list of sensitive words greatly increased the workload of the censorship department, which resulted in the lower quality of censorship."[8]

Second, covert censorship can prevent backlash that comes from censorship, as experiencing censorship can cause individuals to feel angry and escalate their behavior. As *Roberts* (2015) demonstrates, users who find out they have been censored respond with anger, criticism of censors, or by continuing to post off-limits content. Escalation of bad behavior means more work for content moderators which increases the cost of content moderation. Users who are censored but are highly motivated to share a certain type of content may join the "reincarnation party," serially creating new accounts once the previous one has been blocked.

Third, covert censorship increases the appearance that platforms are relatively free and open, which is desirable to users. Sina risks losing users if it gains a reputation

[7] For more information on this event, see a report at ChinaFile here: http://www.webcitation.org/71xsmVFHX

[8] See http://www.webcitation.org/727WJLM9x

for strict censorship, and thus wants to minimize the tangible impact of censorship on users. Experiencing censorship is unpleasant, and Sina Weibo wants users to feel good about using their platform. If users are constantly receiving censorship notices, they may take their business elsewhere. This concern is apparent in the database of leaked censorship logs analyzed below, and is discussed in detail in Chapter 3.

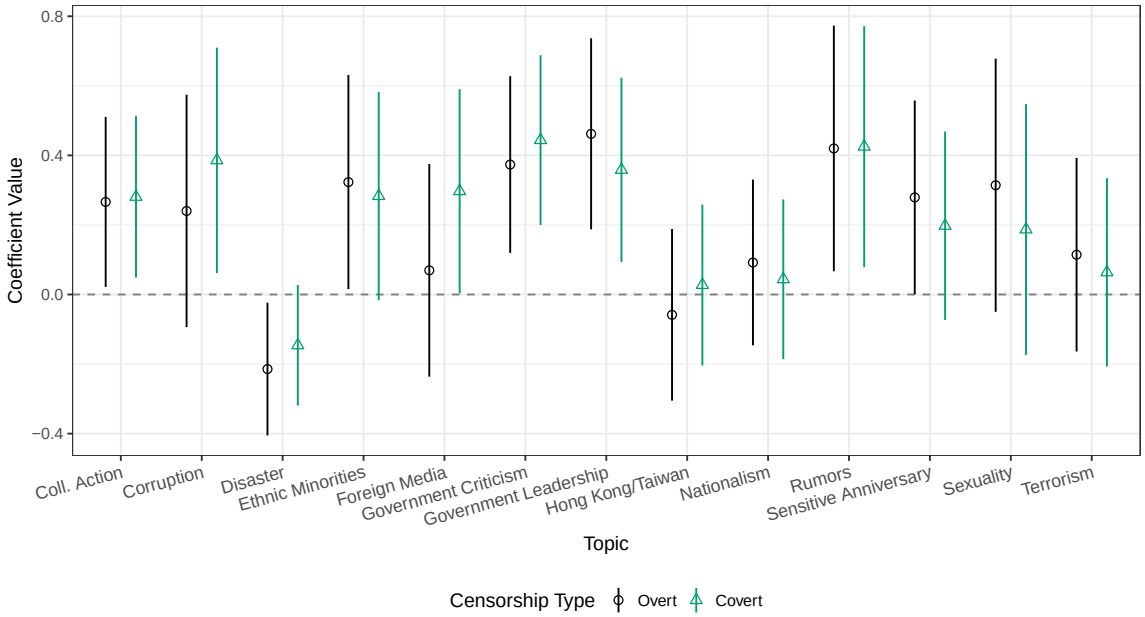
2.5 Empirical Analysis

Using a database of censorship logs from Sina Weibo, I adjudicate between two hypotheses, 1) that covert censorship is selectively used for certain topics of content, punishing discussion of certain types of content through overt censorship, while hiding censorship for content that is less objectionable and 2) that covert censorship's main purpose is to hide censorship from users, regardless of content, since alerting users to censorship can harm a platform's business interests. To test these hypotheses, I manually label the censorship instruction for each log in the leaked log database (the type of censorship Sina Weibo employees are instructed to use for the content in logs), and the topic of content in the logs (for a detailed explanation on coding procedures, please see chapter 4).

I examine the relationship between content and censorship outcomes using multinomial logistic regression comparing the outcome variable, "censorship type," a variable with three unordered levels "do not handle," "covert censorship," and "overt censorship." I include all top-level topic categories as independent variables. I find that certain categories of content are "covertly" or "overtly" censored more often than the baseline category "not handled," namely collective action, government criticism, corruption, ethnic minorities, rumors, and sensitive anniversaries. I also find that the topic "disasters" is more likely to be not handled than handled. This is consistent with a common guideline of openness when reporting on disasters *Zou and Su* (2015); *PRC State Council General Office* (2016). However, there is not a statistical

difference in the coefficients for the "overt" and "covert" censorship levels at the 95% confidence, suggesting that topic does not affect decisions to covertly or overtly censor (see Figure 2.3).

Figure 2.3: Multinomial Logistic Regression Plot of Censorship Type by Topic



Based on the censorship instructions in directives, 78.83% of posts are "secreted." This means that, at least on the Sina Weibo platform, the most common form of censorship is likely to be unobserved. A distribution of the types of censorship in the data can be seen in Table 2.1. In order to identify the purpose of this high level of covert censorship tactics on the Sina Weibo platform, I examined logs where censorship instructions differed by the level of user influence. I find that when a log specifies different censorship methods for users based on their influence, the censorship method suggested for more big/important users is more covert in 80.27% of logs. In Table 2.2, 36.91% of logs with different instructions for "big/important" and "small/ordinary" users recommend deletion for small/ordinary users and secret for big/important users. In 27.9%, covert censorship is recommended for small/ordinary users, and the even

less intrusive "audit" is recommended for big/important users. This is likely due to the relative ease to which big users can infer they have been covertly censored, as they have become used to a baseline level of user interaction with their content. Instead of covertly censoring their content, they simply censor comments on their content and reposts of their content. The texts of logs themselves are also strong evidence supporting the notion that covert censorship tactics are meant to prevent users from finding out they have been censored. In log from 2/18/11, employees were given the following notice about a specific user:

> "Colleagues responsible for audits and user monitoring please be aware: Beijing Zhu Fuxiang has been temporarily put under surveillance and is being audited. He is demolition and rights protection advocate. The Supervision Department says you should not delete his things. If he is too radical, use 'secret.' Don't provoke him."[9]

Logs also frequently mention difficulties arising from the "reincarnation party," which suggest that users finding out they have been censored can have long standing negative impacts on Sina Weibo's censorship division. The reincarnation party consists of users who continuously create new accounts and continue posting "harmful" content after being deleted. Once a user has become a member of this party, they appear to become a continual source of work for Sina Weibo.

2.6 Discussion

How does covert censorship, or more broadly, covert information control such as comment astroturfing (see Chapter 5) increase the effectiveness of information controls in China? Covert censorship hides censorship from users and in doing so obscures the role of the state in the process of censorship. Overt censorship, by contrast, usually requires an explanation for the disappearance of content, often referencing "relevant government regulations." Covert censorship minimizes the anger and backlash that

[9]Original Chinese: 审核负责人监控的同事请注意，北京朱福祥暂时加为负责人监控。他是拆迁维权的，网监 要求：千万不要删除他的东西，如果有过激的都私密。不要招惹他。

Table 2.1: Descriptions of Censorship Types and Other Content Moderation Terms

Name (EN)	**Name (ZH)**	**Type**	**Description**	**Percent of Logs**
Secret	私密	Censorship Method	Hide the post from all from all users but the original poster.	78.83%
Delete	删除	Censorship Method	Remove the post; the original poster and all other users cannot see the post.	16.38%
Don't Handle	不处理	Censorship Method	Take no action on the post or a certain type of post mentioned in the directive.	2.47%
Stop Function-ality	禁止	Censorship Method	Report post/user to Beijing-based "Government Affairs Liaison" for high-level Sina or Beijing Municipal Government officials to deal with personally.	1.89%
Conceal	隐藏	Censorship Method	Hide the post from all users who are not following the original poster.	0.22%
Friend Restrict	友处理	Censorship Method	Hide the post from all users outside of the original poster's friend circle (好友圈). A friend circle is a group of close friends with whom a user chooses to selectively share more personal content.	0.22%
Audit	审核	Surveillance	Add a post or user to a list for employees to manually audit and review comments that are made on that post or user profile.	19.16%
Report Up	上报	Escalation, Repression	The post cannot be retweeted or shared in a private message by anyone.	7.22%

(*Roberts*, 2018) finds results from censorship which is bad for both private internet platforms and the state. By preventing backlash, covert censorship also reduces the occurrence of conflicts between users and the state over the acceptable bounds of discussion. Users who are ambivalent about political issues are less likely to find out through censorship that their opinions are in opposition to the state. This may suppress latent members of the opposition who could be activated by discovery of the state's revealed preferences through censorship.

Covert censorship also may impact user behavior, as objectionable posts will receive no accolades, no likes, and no retweets. It merits further exploration whether depriving users of this positive feedback can lead to changes in behavior or opinion. Many have written on the Chinese state's use of psychological coercion to maintain control of opponents (*Ong*, 2015; *Chen*, 2017; *Cai*, 2008; *Deng and O'Brien*, 2013; *O'Brien and Deng*, 2015). Perhaps covert censorship can serve similar coercive func-

Table 2.2: Logs with Different Instructions for Ordinary and Important Users

Small/Ordinary User Instruction	**Big/Important User Instruction**	**Count**	**Percent**
Delete	Secret	86	36.91
Secret	Audit	65	27.9
Delete	Audit	16	6.87
Secret	Stop	14	6.01
Secret	Delete	12	5.15
Secret	Pass/Allow	6	2.58
Stop	Secret	5	2.15
Other	Other	29	12.45

tions without users attributing blame to the state.

Many of these theories of the impact of covert censorship will be explored in survey experiments that will be fielded in the near future.

2.7 Conclusion

In the data analysis above, I presented evidence that covert censorship is preferred by social media companies because it hide censorship from users, and reduces costly backlash and circumvention efforts that result from users finding out they have been censored. I also outlined many ways in which hiding censorship would be beneficial to the interests of a private company.

CHAPTER III

The Limits of Commercialized Censorship in China

Introduction

Why do scathing criticisms, allegations of government corruption, and content about collective action make it past the censors in China? Past works have theorized that regime strategies or state-society conflicts are the reason for incomplete censorship. While these factors likely contribute to incomplete censorship, I suggest that incomplete censorship results in part from delegation of censorship to private companies which creates a principal-agent problem. Censorship directives are passed through a tangled network of multiple government principals and are delegated to private social media corporations. Government principals and media corporation agents are driven by competing logics: the government logic of information control and the market logic of satisfying user demand for information, respectively.

Using a unique corpus of leaked documents from a social media company, Sina Weibo, I demonstrate that these conflicting logics are the cause of much of censorship's apparent incompleteness. I find that 16% of directives from the government are disobeyed by Sina Weibo and that disobedience is driven by Sina's concerns about censoring more strictly than competitor Tencent. I also find that the fragmentation inherent in the Chinese political system exacerbates this principal agent problem. Fragmentation results in decentralization of censorship enforcement, competition be-

tween government agencies over censorship objectives, and non-uniform distribution of regulatory leverage across the many government agencies in charge of delegating censorship and sanctioning social media companies for non-compliance.

This chapter contributes to our understanding of media control because it shows that market competition impacts information control outcomes, breaking with a large body of works on information control that assume market competition doesn't matter. This chapter complements the work of *Yang* (2013) by showing that market concerns open space for contention in China, but it also emphasizes that market competition is the main driving force behind this opening. Internet businesses push back on government controls in an effort to appear "more free" than competitors, leading to expanded space for contention.

3.1 The Southern Weekend Incident

Each year, the liberal Chinese newspaper Southern Weekend writes a New Years editorial on a theme they would like to characterize the coming year. In January of 2013, they chose the theme "China Dream, Constitutional Dream," stressing the need for progress in the coming year on strengthening the rule of law and protecting rights enshrined in the constitution. After the editorial team submitted their final draft to the Guangzhou Propaganda Department and received no edits, they assumed everything had been approved for publication. The editors and Southern Weekend's readers were surprised when they opened their newspapers and found a fawning paean to the Party in place of the expected boundary-pushing editorial that had become the newspaper's trademark. The article boasted, "we are closer to the Chinese dream than ever before" among other platitudes.

This clumsy reworking of the editorial resulted in a strike by Southern Weekend editorial staff and sizable student protests in major Chinese cities. The strike and protests gained momentum through coordination and discussion on major Chinese so-

cial networking sites such as Sina Weibo[1], despite four strongly worded directives from the Central Propaganda Department and top leadership to cease all such discussion. This apparent incapacity to control perhaps the most threatening type of information online—information with potential to fuel student-led collective action[2]—may come as a surprise to many. The Chinese state is often represented in press and academic writing as a monolith[3], with high capacity to control information flows using the many advanced methods of censorship at its disposal. However, in many circumstances, it seems that intense pressure from the highest levels of the party and government fails to move social media companies like Sina Weibo to act.

[1]According to monthly active user statistics, throughout most of the log data, Sina Weibo was the second largest social network in China, behind Qzone. WeChat, China's most popular social platform is a messaging app that is similar to What'sApp, but with several social, payment, and service features tacked on. Though WeChat's monthly active users surpassed Sina Weibo in Q1 of 2012, it is not a competitor with Sina Weibo in the same way as Tencent Weibo is; both have a similar microblog platform, ostensibly inspired by Twitter. Because Tencent Weibo has been in beta for several years, Tencent does not report MAU numbers in its annual financial reports, however, most measures of active users during the period in which the logs were created put Sina Weibo comfortably in the lead.

[2]The Chinese Communist Party has experienced many student-led movements that have presented clear threats to its grip on power. Students initialized and sustained the Great Proletarian Cultural Revolution, a movement resulting in the dismantling of state and party institutions through arbitrary mass violence. Student—led movements on two separate occasions—in 1976 and in 1989—sparked mass protests which at the time seemed capable of threatening the CCP's monopoly on power.

[3]Much work on censorship assumes that either the government is a unitary actor or that the central government is the main enforcer of censorship. Formal literature, often for the sake of parsimony, defines "the government" or "the autocrat" as the singular actor and practitioner of censorship (*Lorentzen*, 2014; *Guriev and Treisman*, 2015; *Gehlbach and Sonin*, 2014; *Chen and Xu*, 2016; *Egorov et al.*, 2009b) (*King et al.*, 2013) draw inferences from censorship outcomes to measure a singular intention of a single government actor: to allow criticism but censor content with "collective action potential."

3.2 Delegation, Fragmentation, and Agency Loss

During the Southern Weekend editorial incident, the Propaganda Department of Guangzhou[4] and the Central Propaganda Department[5] were involved in attempts to control the spread of information on the event. Despite these urgent attempts to censor all mention of this incident, a novel leaked dataset from 2011-2014 documents that popular social media company Sina Weibo willfully ignored directives to remove content related to the incident on their platform. Sina's calculus was clear. By providing more information about the strike and protests, they could attract information-seeking users away from chief competitor Tencent.[6] One log of company decisions instructs employees to "not be stricter than Tencent," and to hold off on implementing government directives until "urged to block content a second time." When told to delete users, Sina instructed employees to block users temporarily and unblock them the following day "as soon as you receive instructions."[7] The case of the Southern Weekend editorial incident illustrates how government fragmentation and delegation of censorship to private corporations can result in incomplete censorship outcomes. In this case, the Central Propaganda Department, the Guangzhou Propaganda Department, and the Beijing Municipal Government agencies directly

[4]Tuo Zhen, the head of the Guangdong Propaganda Department at the time was concerned about limiting the fallout in response to the editorial incident, especially since he was being blamed by both the public and the Central Government. See this archived analysis for more information: https://web.archive.org/web/20180730135458/http://chinamediaproject.org/2013/01/07/inside-the-southern-weekly-incident/

[5]The original directives can be found archived at the following links: http://www.webcitation.org/71yPYfpyj, http://www.webcitation.org/71yPahJIQ, and http://www.webcitation.org/71yPeE80B

[6]Tencent is the largest ICP company in China. It owns WeChat, Tencent Weibo, and QZone, three of China's most popular social platforms. It is Sina's most direct competitor.

[7]Full text of log from 1/5/2013: "There is a lot of related content on Tencent. After we reported [their lack of implementation] to the Network Management Office, Tencent implemented [censorship of the content]. Currently [Sina] Weibo is partially carrying out instructions to block content. First prevent retweets on content flagged [by the Network Management Office]. When urged to block content a second time, fully implement directives. With respect to banning users, for the meantime do not implement [user bans]... handle users relative to Tencent's [level of] implementation. We should not be stricter than Tencent. Today maintain user blocks, tomorrow as soon as you receive instructions release the block."

involved in managing Sina Weibo had different objectives and were unable to adequately monitor and sanction the company. By delegating censorship to Sina Weibo, concerns over user retention and competitiveness became a factor in the company's decision to comply with directives. By Sina's calculations, the cost of flouting government directives outweighed the benefits of increased user engagement. In this section I outline this theory in detail. In subsequent sections, I present evidence in support of this theory from censorship logs like the one mentioned above and large databases that measure censorship outcomes on the Sina Weibo platform.

3.2.1 Corporate Delegation Leads to Agency Loss

Private companies play a crucial role in the process of censorship, as they bear the ultimate responsibility for removing content from their platforms. Despite their central role, private companies are too often missing from models of information manipulation. In China, censorship is delegated and regulated through internet content providers (ICP) licenses. ICP licenses are necessary to operate an internet business in China. These licenses can be revoked or suspended if ICPs do not comply with government directives.

Sina Weibo and many ICPs like it have had to balance user and shareholder demand with regulatory pressures from government agencies since they were first founded. In 2014, when Sina Weibo became listed on the NASDAQ stock exchange, it made these concerns explicit. In its regulatory filing with the SEC, it included "regulation and censorship of information disseminated over the internet in China" as a major risk that could affect its share price. In the filing, Sina Weibo also noted that censorship "may adversely affect our user experience and reduce users' engagement and activities on our platform as well as adversely affect our ability to attract new users to our platform."[8] Since Sina's listing on NASDAQ the company has been

[8]Full article accessible here: https://web.archive.org/web/20180812134506/

fined several times by Chinese government agencies for failing to meet censorship regulations. Sina's user engagement and activity has declined in recent years, in part due to the rising popularity of WeChat, a chat-based social networking site like WhatsApp. Some analysts, however, have attributed this decline to increasing perceptions of Sina Weibo as a heavily-censored platform.[9]

I argue that "incomplete" censorship is largely a result of a clash between ICPs responding to popular demand for information and government actors responding to pressures to maintain social stability and protect their position within government. The process of delegating to private internet companies creates a principal-agent problem. Government principals delegate censorship to private internet companies through directives: verbal or written instructions providing details about how these private internet companies are supposed to handle certain kinds of objectionable content. Private internet companies then decide if and how they will comply. Because there is often a misalignment of preferences between the government principal and the ICP agent, private internet companies may ignore the directive or partially implement it. Because of this disobedience, the delegating principal suffers agency loss: agents acting against their principals' interests.

Competition between private internet companies further exacerbates this problem. In China, each government principal delegates censorship to several private internet companies, meaning that there are multiple agents involved in China's system of censorship. *Rundlett and Svolik* (2016) have shown that when a principal has multiple agents, principals suffer agency loss. Private internet companies benefit from having compelling information on their platforms. However, compelling content may also be considered "harmful information" to the authorities. Highly motivated users who are seeking this information often hop from platform to platform when content is censored (*Roberts*, 2017). By censoring less, private internet companies can attract

[9]See: http://www.webcitation.org/72Kwc4BJ7

these users to their platform and away from their competitors. This competition over users results in a race to the bottom as each company strategically tries to skirt directives more than competitors. Companies seek to jointly minimize the cost of censorship (lost user retention and engagement) and the cost of non-compliance with government directives.

Private internet companies deliberately shirk, lack capacity to implement directives, or some combination of the two. Private internet companies can take advantage of hidden information about their technical capabilities and budgets to invest as little as possible in developing high-performance censorship systems. They can also shirk, as their effort censoring content is not easy to measure or observe. Private internet companies take advantage of hidden actions and hidden information to skirt directives when they anticipate that the benefits outweigh the costs.

3.2.2 Bureaucratic Fragmentation Leads to Agency Loss

Fragmentation of China's political system further complicates censorship delegation. While on paper the Chinese political system is rigidly hierarchical, in practice, contestation between state and party organs, and bureaucracies with overlapping policy domains is common throughout the system. The findings of this analysis confirm earlier suppositions about the fragmentation inherent in the Chinese political system.[10] This fragmented system, which *Oksenberg and Lieberthal* (1988) coined "fragmented authoritarianism" is characterized by de facto veto power of local governments when implementing central policies and the tangled lines of authority in China's vast bureaucracy. I argue that incomplete censorship is in part due to the fragmentation of China's political system. This fragmentation leads to agency loss—disobedience of, or incapacity to implement government directives—due to two major attributes of China's system of censorship: common agency and local bias.

[10]See *O'Brien* (1994); *Montinola et al.* (1995); *Lieberthal* (1995); *Jin et al.* (2005); *Zheng* (2007); *Stern and O'Brien* (2012); *Mertha* (2009)

First, common agency[11] of private internet companies gives them discretion about which agency's directives to follow. Many bureaucracies in China have the authority to delegate censorship to private internet companies. This makes monitoring and sanctioning of private internet companies difficult because each individual bureaucracy must rely on their own limited resources in order to monitor compliance after they have issued their directive. Private internet companies receive directives from multiple principals (sometimes referred to as "common agency"), so they often have discretion about which directives to follow, especially with principals do not have uniform preferences (*Calvert et al.*, 1989), as is often the case in China. Because of this, it is not uncommon to observe inconsistencies in the way content is censored across private internet companies due to this discretion.[12]

Second, the power to enforce directives is locally biased, i.e. power is concentrated in the locality of an ICP's headquarters. Local agencies where media companies are headquartered have more control over what gets censored as they have access to the most proximate and effective levers of power to enforce compliance with directives. This leads to agency loss because the central government and local governments outside of the ICP's jurisdiction do not have the same implements of enforcement and are easier to ignore. Because private internet companies are constantly weighing the cost of non-enforcement and expected revenue from satisfying user demand, private internet companies will pay outsized attention to local directives, despite being national platforms. Without a centralization of the power to issue and enforce directives, private internet companies can selectively ignore directives from outside of

[11]Many government agencies have the power to issue directives to single private internet companies.

[12]For example, there is significant correlation among censored keyword lists in Chinese game chatrooms when they are created by the same parent company or developer, but very weak correlation among lists within the same Chinese provincial or city jurisdiction (*Knockel et al.*, 2017). Previous work has found inconsistencies in the implementation of censorship across platforms and companies operating in China, including search engines (*Villeneuve*, 2008), blogging services (*MacKinnon*, 2009), chat apps (*Crandall et al.*, 2013), live streaming (*Crete-Nishihata et al.*, 2016), and mobile games (*Knockel et al.*, 2017), which suggests companies have flexibility and discretion when interpreting and implementing censorship directives.

their jurisdiction.

The complexities of information control in China's fragmented system have been well-studied as they relate to traditional media. In a study of newspapers, *Stockmann* (2013) finds that during the reform of China's media system, fragmentation posed a challenge to state monitoring and sanctioning of commercialized newspapers. She argues that factional affiliations, rank of a media company's sponsoring agency, and geographical jurisdiction defined a "discursive space," giving media companies greater freedom to report more critically about opposing factions, lower ranked agencies, or other geographical regions. *Mertha* (2009) finds that policy entrepreneurs make use of the media to lobby and appeal to various fragmented interests across China's bureaucracy. Others have noted the many conflicts between central and local governments, party and state organs, propaganda departments in different localities, and media organizations and regulators (*Brady*, 2009; *Lynch*, 1999; *Shambaugh*, 2007).

3.3 Data and Methods

To test the above theoretical claims, that fragmentation and corporate delegation result in outcomes that deviate from the intentions of delegating government principals, researchers face a number of challenges. Despite the vital role private internet companies play in China's system of censorship, there has been a dearth of data on them and very little scholarship dedicated to them. Much of the available data used to study Chinese censorship consists of only content and censorship outcomes. These data do not provide any information about the interactions between individuals in government and between private internet companies and government actors that determine what is and is not censored. Due to limitations of available data, assumptions that censorship outcomes can serve as a measure of government intent are common, despite the central role non-government actors play in the process.

3.3.1 Log Data

To address these shortcomings, I have created a custom dataset of censorship logs—notes taken in the process of censorship at Sina Weibo, one of China's most popular social networking sites. This dataset is the first of its kind to capture the entire process of Chinese censorship, from a government directive to a private internet company, to that private internet company's decision on how or whether to comply. With the help of research assistants, I have coded these logs by content, the bureaucracy issuing directives to Sina Weibo, and whether or not Sina Weibo implemented these directives. Log data adds empirical support to the theoretical claims detailed earlier in the following ways:

First, because these logs measure disobedience, they can provide insights into the relative power of bureaucracies to delegate censorship. By examining the variance in Sina's rates of compliance with directives across different bureaucracies, we can observe how fragmentation leads to agency loss. If fragmentation results in agency loss, rates of compliance with directives should be correlated with a bureaucracy's power to monitor and sanction Sina Weibo.

Second, logs include direct and indirect information about the reasons behind non-compliance with directives. Occasionally managers explain their decisions to comply or disobey government directives. If corporate delegation leads to agency loss due to concerns about competitors and the adverse impact of censorship on user experience, Sina Weibo managers should express these concerns in the logs. Other less direct evidence of profit motivations can be observed in how thoroughly Sina censors content that is harmful to its business interests and content that is harmful to government interests. By comparing the censorship rates of Sina-related news events and censorship on behalf of a government actor, these logs can add support to claims that Sina Weibo factors market concerns into its censorship efforts. Comparing censorship rates of content that is harmful to Sina's business interests and content

that is harmful to government interests can also demonstrate that agency loss is not simply due to a lack of capacity, but rather due to deliberate and strategic shirking.

3.3.2 Event Data

Users seek information at different rates depending on the type of event, and more user attention is usually more threatening to the government. Because Sina is concerned with audience demand, they may increase or decrease their censorship efforts in response to audience interest and, by proxy, sensitivity. To see how censorship implementation varied by audience demand for information for sensitive events, I used data from annual "blue books," policy briefs written for government cadres, on "social opinion and emergency management" compiled by the Institute for Public Opinion Research of Shanghai Jiao Tong University and published by the Chinese Academy of Social Sciences. These blue books collate reports written by academics, government officials, and policy experts and offer suggestions for policy changes and improvements in the coming year. Using these events as a sample ensures that I am analyzing events that public opinion experts found noteworthy, and where "public opinion supervision" (舆论监督) and/or "public opinion guidance" (舆论引导) was potentially necessary. These reports include 300 events (60 events for each year from 2010-2014) ranked by an index of search volume across several services. Each year's cases are selected from the top 1200 events with the highest search volume index[13] based on their designation as a "public opinion emergency" that would be salient to opinion and thought workers, the intended audience of these "blue books."

[13]The search volume index is calculated as the average of news search volume, social networking platform search volume, blog search volume, Sina Weibo search volume, video search volume, and WeChat Public Accounts search volume.

3.3.3 Censorship Outcomes Data

For each of the events above, I measure how thoroughly Sina carried out censorship on its platform by searching large datasets of historical Sina Weibo posts for the exact content targeted for censorship in logs. For this analysis I use two datasets that measure censorship on the Sina Weibo platform: Free Weibo and Weiboscope. The Free Weibo dataset includes 47 million weibo posts spanning from 2009-2018 and the Weiboscope dataset includes 226 million posts from only 2012.[14] Using these data I compare the instructions in the logs to actual censorship outcomes at the event level.

To compare log content to actual Weibo post content, I first had research assistants manually identify logs relevant to each of the 320 events from "blue books" by searching within a month window on either side of the event date. For each relevant log, research assistants then extracted the full content text from logs, stripping away instructions on censorship and government directives that usually go along with log text. After content from relevant logs was extracted, I searched large databases of censorship outcomes for exact or near-exact text matches to each individual log's content.

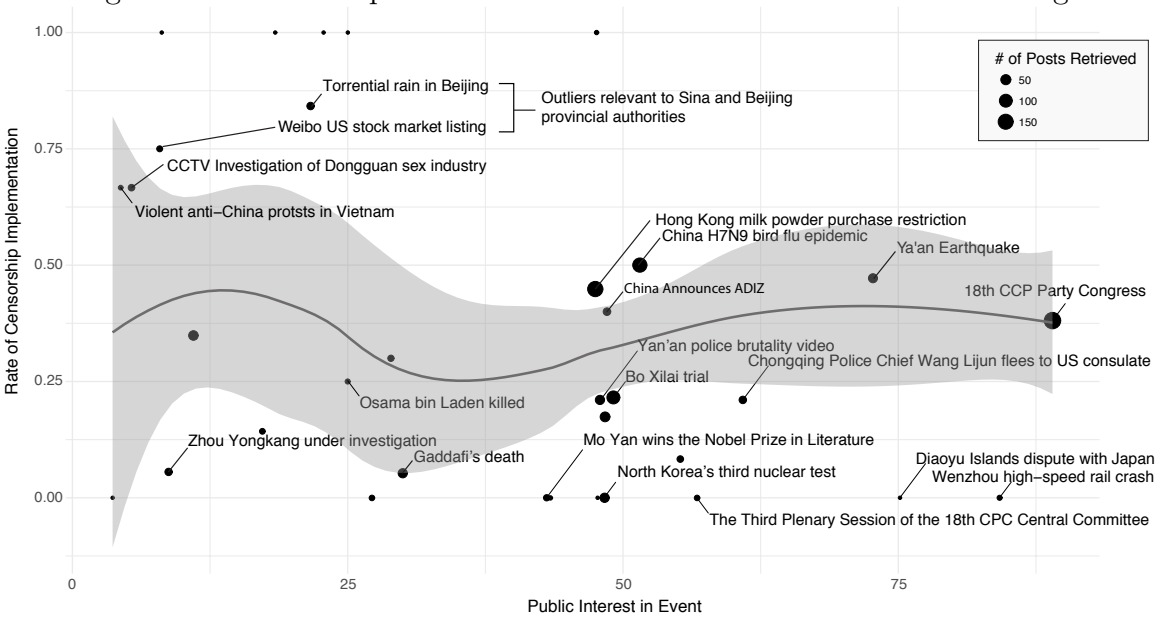
Running several searches of 273 million Weibo posts is not a trivial task. To identify exact or near-exact matches to posts mentioned in the logs, I needed to build a rudimentary search engine, building an inverted index of each of the Weibo posts in these large databases. For each log, I then scored the relevance of each post in the two large databases of Weibo Posts using the Okapi BM25 ranking algorithm. The Okapi BM25 algorithm was the gold standard for search engine ranking prior to recent advances in deep learning and was the core technology behind the Bing search engine for several years. With each post in these databases ranked according to relevance to the text extracted from the logs, research assistants then identified posts from each

[14]Both datasets are constructed by recording a post soon after it has been created and later querying for that post at fixed intervals to see if the post has been censored or deleted.

set of search results that exactly or almost exactly matched the post targeted in logs. Results were considered near-exact matches if they contained the entire query text with either small editions, or a a few words replaced with synonyms, homophones, or homoglyphs. Using the final dataset of retrieved matches, I estimated the proportion of posts that were actually censored by Sina Weibo for each event. The results of this procedure are visualized in Figure 3.1.

3.4 Empirical Implications and Results

Figure 3.1: Censorship Rate of Retrieved Weibo Posts Mentioned in Logs



Each point on the figure represents a event where posts were retrieved and its size is proportional to how many posts were retrieved. The y-axis, "rate of censorship implementation" is the number of retrieved posts that were censored over the total number of posts retrieved. The x-axis, "public interest in the event" is the search volume index gathered from blue books. The regression line and confidence intervals are from a weighted LOESS model. The average rate of censorship implementation is .38.

3.4.1 Corporate Delegation Leads to Agency Loss

I argue that incomplete censorship is due in part to a clash between the market logic driving the behavior of private companies and the logic of control that guides governments. In other terms, market demand for information sometimes clashes with the informational preferences of government actors. In this section, I argue that corporate delegation leads to agency loss for three main reasons. First, market competition incentivizes Sina to prefer to censor less than its competitors, leading to a race to the bottom. Second, Sina Weibo is technically limited in its ability to comply with government directives, resulting in agency loss. Third, Sina Weibo deliberately shirks in order to minimize the cost of censorship and to minimize impact of censorship on the usability of its platform.

The logs provide many examples of Sina Weibo's concerns about censoring more than competitor, Tencent Weibo. In one instance Sina Weibo was instructed to delete all content related to a murder case that had garnered much public interest. According to logs, Sina Weibo monitored the compliance of Tencent Weibo, alerted the Supervision Department to their non-compliance, and delayed implementation of directives until they could be assured that Tencent had also complied with directives.[15] In another example, the Network Management Office ordered Sina Weibo to remove a popular account on their service. After seeing that Tencent had not complied with removing the same user on their platform, Sina Weibo drafted a response to the Network Management Office: "Since Tencent hasn't deleted the account, we are unable to delete the account at this point in time." In several other cases, Sina waited until Tencent complied to implement directives to delete user accounts. During several

[15]Log entry from December 26, 2011 reads: Supervision Department demands to eliminate related content to the Henan Shenqiu murder case that caused the death of four children and injury of one child. Currently, we told Supervision Department that there are many related posts on Tencent Weibo and asked Tan Chao to negotiate. They have not yet responded. Previously, we censored any searches of the incident but didn't eliminate everything. Currently, the parameter is to process anything attacking the party or government policy. If there are other situations, report immediately.

sensitive events such as the annual Spring Festival Gala—a Chinese propaganda variety show and most-watched television show on the planet[16]—Sina Weibo instructed employees to avoid censoring more comments on the official television broadcaster's account than competitor Tencent. Sina also went to great lengths to limit positive news about Tencent on its platform. When news broke about a mass purge of bots on Tencent, Sina Weibo employees were instructed to prevent retweets of posts that praised Tencent's actions. Similarly, when Xi Jinping visited Tencent, Sina Weibo employees were instructed to prevent retweets of all related posts. Conservatively, these patterns in logs suggest that Sina Weibo's concerns about remaining competitive with Tencent drove many of their decisions to comply with directives. Preferences to censor slightly less than competitors, and the low level of implementation of directives is consistent with a race to the bottom.

Though we observe Sina deliberately disobeying directives in logs, in many instances, it seems that Sina's poor censorship performance is due to a lack of capacity. In Sina's IPO documents they claimed, "although we attempt to monitor the content posted by users on our platform, we are not able to effectively control or restrict content generated or placed on our platform by our users." This statement appears to be at least partially accurate. In some logs, Sina Weibo appears overwhelmed. During a major collective action incident in 2012 involving mass protest and several self-immolations in Tibet, Sina Weibo struggled to keep up with a large magnitude of takedown requests about Tibet. At one point, Sina Weibo employees sent an SOS to all department heads as they struggled to mobilize enough employees to meet increasing censorship demands.[17] Even in less dire circumstances, Sina Weibo appears

[16]Viewership of the New Year's Gala is around 700 million. In 2012, Guinness World Records gave the show a audience of 498.7 million and named it the, "Most Watched National Network TV Broadcast." See http://www.webcitation.org/71y5Jcwx5.

[17]Log entry from February 1, 2012 reads: Recently there have been a lot of demands to delete posts about Tibet!! There is no way we can add this many banned keywords; if we do so, the rate of posts needing investigating will increase too much. We have already added several keywords related to Tibetan independence, Communist Party [policy in Tibet], [police] killings, but the number of keywords keeps increasing... I don't know what to say!! Every department head, please disseminate!

to perform relatively poorly. In Figure 3.1, we see that Sina Weibo usually does not perfectly follow through with the decisions to comply with censorship directives. The average rate of censorship of Sina Weibo posts from the Free Weibo and Weiboscope databases that exactly match posts in the logs is .38. This low rate of censorship implementation may be for one of two reasons: unintentional low capacity, or deliberate shirking.

To test whether Sina's poor performance was exclusively due to low censorship capacity, I identified all logs that referenced news events about Sina Weibo that were potentially damaging to the company's reputation or bottom line. All posts mentioning Sina had already been labeled by research assistants, making the search easier. In total I identified three such news events in the logs: 1) discussions of Sina's IPO, 2) discussions of new research measuring the speed of Sina's censorship, and 3) Reuters interviews with former Sina Weibo employees who worked as content censors. To measure how well Sina followed through on these decisions to censor content, I retrieved exact or near-exact matches of content mentioned in the logs from the Free Weibo database using methods described in detail above. In total I retrieved 33 posts, 24 of which were censored, an implementation rate of .73. This rate of implementation is significantly higher than the Blue Book rate mentioned earlier (.38) according to a 2-sample chi-squared test of equality of proportions at the 95% confidence level. This evidence suggests that Sina Weibo has the capability to more thoroughly censor content in response to government directives, but chooses not to. I find that the low rate of actual censorship is likely at least somewhat deliberate.

These quantitative measures are consistent with Sina's general instructions about how intensely to censor content. Employees are at times instructed to deliberately shirk and to obstruct the process of censorship by delaying directives and prolonging bargaining between delegating government agencies and Sina. In the days preceding

This is extremely urgent!!!

the Southern Weekend editorial incident mentioned in the beginning of this chapter, log documents begin with a general notice urged employees to "negotiate as much as possible" and "defer implementation of censorship requests from the Supervision Department and Internet Management Office," Sina Weibo's two chief regulators. The notice went on to instruct employees to "not process too many user posts" and to "not be too stringent." Just three months earlier, a similar general notice had suggested the opposite. Employees were urged to "tighten" their "control measures" and to "resolutely eliminate all posts relating to negative incidents, rumors about the Politburo Standing Committee and their families, collective action content, coups, power struggles related to the 18th CCP Party Congress, Ling Jihua, Central Public Security Bureau, Bo Xilai, etc." But even this strongly worded notice instructed employees to refrain from immediately implementing "especially unreasonable directives."

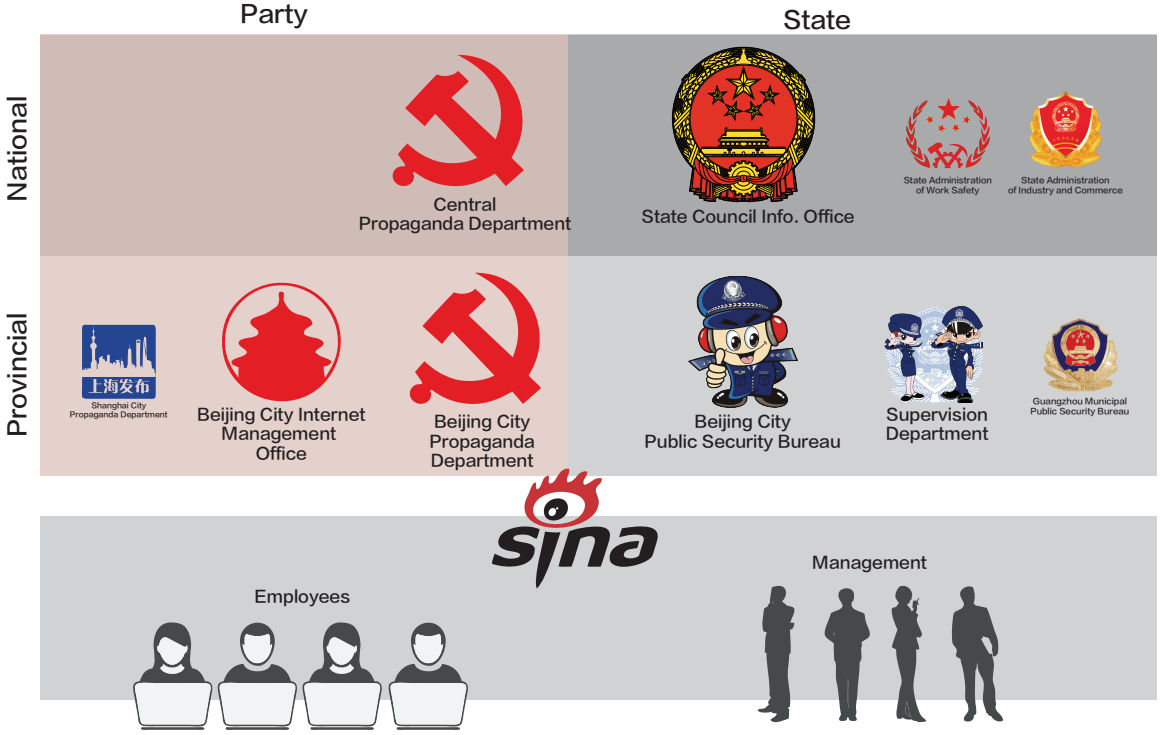
Finally, as is the case in any workplace, part of Sina Weibo's lack of capacity to censor content comes from agency loss due to rogue or incompetent employees. Of course, the individual who leaked this entire cache of documents to the press was a Sina Weibo employee.

3.4.2 Bureaucratic Fragmentation Leads to Agency Loss

Leaked censorship logs from Sina Weibo do not depict a system of censorship that is hierarchical, centralized and efficient. Rather, many bureaucracies appear to have varying degrees of authority, do not appear to coordinate their censorship directives, and appear to at times be in conflict with one another. Logs document 3 main bureaucracies in charge of censorship at Sina Weibo, and a long tail of other bureaucracies with the power to issue directives to the company (see Figure 3.2). To measure the relationship between fragmentation and agency loss, I calculate rates of disobedience across multiple bureaucracies, and confirm that these patterns match theoretical expectations. I then argue for two mechanisms behind this relationship.

First, common agency of private internet companies gives them discretion about which agency's directives to follow. Second, the power to enforce directives is locally biased, i.e. power is concentrated in the locality of an ICP's headquarters.

Figure 3.2: Agencies and Individuals Influencing Censorship on Sina Weibo



If bureaucratic fragmentation leads to agency loss, we should expect to see differences in the responsiveness of Sina Weibo to certain bureaucracies. Three bureaucracies are most commonly called: The Beijing Municipal Internet Propaganda Management Office (Internet Management Office), the Public Information and Internet Safety Supervision Department of the Beijing Public Security Bureau (Supervision Department), and the State Council Information Office (SCIO).[18] Personnel and budgets of the Internet Management Office and the Supervision Department are controlled by the Beijing Municipal Government. In a handful of cases, the State

[18]The logs confirm much of the basic bureaucratic structures that monitor and direct censorship at Sina Weibo as identified in *Cairns* (2016a).

Council Information Office delegates directly to Sina Weibo, but most of the time it issues instructions to the Supervision Department and Internet Management Office to delegate to private internet companies. Though it has informal authority to delegate to these agencies, it does not have leverage over budgets and personnel. Because Beijing Municipal Government agencies are directly responsible for monitoring and sanctioning Sina, they should be obeyed at higher rates than the State Council Information Office. Beijing Municipal Government has strong incentives to keep Sina in check, as they will ultimately bear responsibility for failure to manage Sina when they are evaluated for promotion at the end of their terms. In the logs, I find that overall, 16% of all directives are disobeyed. Beijing municipal bureaucracies, the "Network Management Office" and the "Supervision Department" are disobeyed at rates of 15% and 17% respectively. By contrast, the SCIO is disobeyed 20% of the time. Though these measures of disobedience do indicate that the SCIO is disobeyed more often than Beijing bureaucracies, there are not enough SCIO directives to distinguish a statistically significant difference in the two proportions. To address this limitation, I examine the data for evidence for two theorized mechanisms linking bureaucratic fragmentation to agency loss: common agency and local bias.

If common agency of private internet companies results in agency loss, we might expect to see instances in the data where agencies delegating censorship send different directives to Sina Weibo, and where Sina Weibo chooses to implement the more lenient or less-specified directive. This is a very hard test to pass because the logs are usually not detailed enough to show discrepancies between directives from two agencies. Directives are usually not copied verbatim and are summarized, often in shorthand. Despite this, there are a handful of logs where we observe behavior consistent with common agency leading to agency loss. One such log involves large-scale protests in the city of Shifang in July of 2012 over a copper plant local residents believed was causing health problems. The Supervision Department, a Beijing mu-

nicipal state public security organ directed Sina Weibo to remove all collective action content from their site. At the same time, the Internet Management Office, a Beijing municipal party propaganda organ directed Sina to remove a list of specific posts.[19] The latter order was easier to implement and was unlikely to completely shut down discussion of the event on Sina Weibo. Sina opted to ignore the first order and implement the latter order.

This evidence has its limitations. These handful of accounts confirm that common agency resulted in some agency loss, but it is impossible with these data to determine the magnitude of this agency loss. Previous work on delegation, however, has shown that in situations where there are multiple principals, agent discretion results in higher agency loss than if there was only one delegating principal (*Calvert et al.*, 1989).

If local bias results in agency loss, we can expect to observe two main things from the log data. First, it should be uncommon for non-Beijing Municipal Government agencies to appear in directives because non-Beijing bureaucracies—especially non-Beijing local governments— should have very little power to sanction Sina for non-compliance with directives. In the case of Sina, the company falls under the jurisdiction of the provincial-level municipal government of Beijing, as Sina is headquartered in Beijing. To measure the distribution of directives sources, two research assistants coded the source bureaucracy for all 8,427 unique censorship logs according to specific instructions about what constituted a bureaucratic source. I manually checked all of their labels and searched the database for any logs they may have missed. The resulting distribution of bureaucracies can be seen in Table 3.1. Overall, 96.6 percent of directives come from Beijing Municipal bureaucracies. While other provincial-level bureaucracies can send directives to Sina, only two out of 611 specified

[19]A log from July 3, 2012 reads: "Regarding the Shifang incident, the Supervision Department is currently demanding that we eliminate any inflammatory and mobilizing content. The Internet Management Office has no clear directives but has sent over a list of individual posts to process, ordinary requests like these should be followed."

Table 3.1: Distribution of Bureaucracies Issuing Directives

Name	**Administrative Rank**	**Beijing**	**Directive Count**	**Percent of Directives**
Beijing Municipal Internet Management Office	Provincial-level Municipality	Yes	310	50.65
Beijing Municipal Supervision Department	Provincial-level Municipality	Yes	269	43.95
Beijing Municipal Internet Police	Provincial-level Municipality	Yes	8	1.31
Beijing Municipal Bureau of Radio and Television	Provincial-level Municipality	Yes	3	0.49
Shanghai Municipal Propaganda Department	Provincial-level Municipality	No	1	0.16
Guangzhou Municipal Supervision Department	Provincial-level Municipality	No	1	0.16
State Council Information Office	National	No	16	2.61
Ministry of Public Security of the Central People's Government	National	No	2	0.33
Central Military Commission of the People's Liberation Army	National	No	1	0.16
State Administration for Industry and Commerce of the People's Republic of China	National	No	1	0.16

directive sources in the logs are non-Beijing Municipal bureaucracies.[20]

If local bias results in agency loss, we should also see greater responsiveness to the demands of the Beijing Municipal Government than other bureaucracies, particularly the Beijing Municipal Public Security Bureau and the Beijing Municipal Propaganda Department. To test this, research assistants coded all instances in the logs where Sina Weibo employees were directed defend government Weibo accounts, monitoring and deleting comments that were offensive to that particular agency. Of the accounts Sina Weibo protected, 46% were Beijing Municipal Government accounts, 46% were national-level government accounts, and 8% were non-Beijing provincial-level government accounts. As Sina Weibo users are not overwhelmingly concentrated in Beijing, this suggests that decisions to censor content are locally biased.

[20]One log includes a directive from the Shanghai Municipal Propaganda Department and another from the Guangzhou Municipal Public Security Bureau.

Sina also appears to censor more efficiently when an event is salient to the Beijing Municipal Government. In the aftermath of a flash flood in Beijing, citizens tried to organize vigils for victims of the flood. Many logs related to this event indicate the Supervision Department and the Internet Management Office's keen interest in thoroughly removing this mobilizing content from Sina Weibo.[21] There was not much interest in this particular event according to the blue book measures (in blue books, the degree of interest in events is a proxy for their politial sensitivity to the national government), and it appears to be an outlier in Figure 3.1. Censorship of this event was much more efficiently implemented than other blue book events at all levels of interest in the event (see Figure 3.1). This is consistent with the theory that local bias due to fragmentation leads to agency loss in the process of censorship delegation.

3.5 Conclusion

Leaked censorship logs from Sina Weibo provide an intimate look into the conflicting informational preferences of the Chinese government and private internet companies. They depict a system of multiple principals and multiple agents and a tangled web of competing informational objectives. The logs show that the outcome of censorship involves many actors and does not necessarily reflect a unified government strategy. Rather, government fragmentation and delegation to several corporate actors results in a system where the end result of censorship is generated by the aggregated and contested preferences of central leadership, subnational governments, and subnational elites passed through a final layer of distortion: media corporations. The logs document outright disobedience of directives, even in highly sensitive situations, and show how Sina Weibo strategically disobeys directives in order to gain a edge over

[21] A log from July 28, 2012 reads: "The Supervision Department requested we ramp up handling and elimination of posts inciting and mobilizing netizens to hold vigils for victims of the Beijing torrential rain disaster. A mobilizing post was discovered tonight and reported to the Supervision Department and Internet Management Office."

competitor Tencent Weibo. While much of the academic literature and media depicts China's censorship apparatus as swift, centralized, focused, and sophisticated, the system is often slow, fragmented, contentious, and low-tech, making censorship orders difficult to enforce and giving social media companies a great deal of discretion over what citizens do and do not see. Delegated censorship to private companies results in significant agency loss that is then further compounded by political fragmentation.

3.6 Developments Since 2014

In the last few months of log data, China's information control institutions underwent significant reforms. These reforms seemed aimed at addressing the problem of fragmentation of China's system of information control. In 2014, the Cyberspace Administration of China (CAC) assumed its role as China's chief regulator of cyberspace. The CAC is a joint party and state organ that houses the SIIO (which is a continuation of the SCIO[22]), and the General Office of the Central Leading Group for Internet Security and Informatization which reports directly to the Central Committee of the Chinese Communist Party. The creation of the CAC gave party and state organs unambiguous authority over provincial and municipal bureaucracies regulating private internet companies such as Sina Weibo.[23] It is unclear from available data whether or not these reforms succeeded in reducing agency loss resulting from political fragmentation. Since these reforms, however, Sina and Tencent still appear to resist and defy regulations.

In 2015, the CAC threatened to shut down Sina Weibo due to insufficient censorship.[24] In 2017, the CAC imposed "maximum fines" on Sina Weibo, Tencent, and Baidu for "failing to fulfill their management duties and violating China's Cyber Se-

[22]Technically, the SCIO was referred to as the SIIO after 2011. In late 2011 reforms, the SCIO's rank was elevated so that it directly reported to the State Council.

[23]See full report on the CAC here.

[24]See Wall Street Journal article here.

curity Law."[25] In 2018 Sina was ordered by the CAC to suspend "key portals such as its hot search site and portal on celebrities and their personal lives for a week" due to its violation of "relevant internet laws and regulations and spread illegal information."[26] In 2018 the CAC suspended popular social networking site, Zhihu for one week for "lax supervision and the spread of illegal information."[27] In late 2018 the CAC ordered the suspension of news aggregators for several weeks due to "illegal" information sharing. Sina Weibo has responded by expanding its efforts to censor content through crowdsourcing and gamification, offering iPads to the best "Weibo Supervisors," users who volunteer their efforts to help Sina clean up harmful content.[28]

These developments are interesting. While Sina appears to increase efforts to police harmful information, it does so through crowd-sourcers, in an attempt to cut costs. In recent state media, the authorities noted that despite these increased efforts, they "are not fully performing their duties,"[29] indicating that recentralization has yet to solve problems of delegation.

3.7 Beyond China

Tensions and alliances between corporations and government actors, as well as government delegation to corporations, are relevant far beyond the case of censorship in China. The fraught alliance between social media companies and governments is representative of a greater trend beyond the Chinese context, in authoritarian and democratic politics alike. Corporations, especially in the realm of surveillance and data analytics are operating on the behalf of governments to spy on citizens, en-

[25]See Global Times article here: http://www.webcitation.org/71y4srDVO, and see a The Diplomat article here: http://www.webcitation.org/71y4v72U0.
[26]See http://www.webcitation.org/71y4wnS1A.
[27]See http://www.webcitation.org/71y52ogK6.
[28]See http://www.webcitation.org/71y53kuQE.
[29]See http://www.webcitation.org/71y55KoD9.

force copyright laws, censor information, and repress government opponents. These alliances represent a "broader trend of neoliberal restructuring, in which political authority and decision-making power are taken out of the public realm and transferred to private environments, often underpinned by commercial and market logics" (*Crouch*, 2004; *Hintz*, 2016). Since the 2016 elections, it became increasingly clear how difficult it would be to hold social media companies accountable for their actions. Behaviors that are in the public interest, such as cleaning up bot accounts, preventing the spread of fake news, identifying and disrupting foreign influence campaigns, monitoring hate speech, preventing the spread of violence, and removing bad actors are often in conflict with fundamental profit motivations and concerns about competitiveness. It took years for Twitter to take any meaningful action on bot accounts, and they did so only under extreme public pressure, due to fears of a hit to monthly active user statistics that would reduce its stock price.[30]

As the Snowden leak revealed, the PRISM program gave the NSA authority to request that Microsoft, Facebook, Apple, and Google provide data matching keywords approved by a U.S. Foreign Intelligence Surveillance (FISA) Court ruling. At large tech companies, requests for information controls are released publicly, showing a tension between corporate and consumer preferences and states' logic of social control (*Tanash et al.*, 2015). Private companies often refuse requests out of concerns for their users' preferences, their profitability, and their reputation. An example of these conflicts is Apple's refusal to cooperate with FBI requests following the San Bernardino terrorist attack. The FBI sought to compel Apple to break their own encryption so that the FBI could obtain information on one of the suspect's iPhones. Obeying such a request, Apple said, would "threaten the security of our customers," which Apple has trumpeted as an advantage over competitor Google. In 2014, they boasted, "unlike our competitors, Apple cannot bypass your passcode and therefore

[30]See http://www.webcitation.org/71y56kNiL.

cannot access [customer] data." Despite public statements about Apple's company values, their refusal is likely in part driven by concerns about profits and competitiveness. Like Sina, Apple is not keen on giving into government demands that would put it at a competitive disadvantage.

The relevance of tensions and partnerships between state and corporate actors is also indicative of trends in Chinese reform that not only includes corporate actors in the policymaking process, but uses corporations as labs to create structures and institutions that can later be subsumed into the state. For example, the very process of censorship described in this chapter has already been partially subsumed into the state. Employees working in editorial functions at internet content producers (ICPs) in China can no longer remain on private payrolls. This means that according to Chinese law, the employees who censor content can no longer be employees of Sina Weibo, but now must be employees of a government agency such as the Network Management Office or the Supervision Department. Whether or not this solves principal agent problems identified in the censorship system has yet to be seen, but control over the internet has been tightening at a rapid pace over the last few years.

Other such plans to subsume institutions borne out of corporations seem on the horizon. China is now relying on the infrastructure of Alibaba's Alipay to serve as the technical back-end to a national "social credit system" which merges social and financial data to give users a score that can selectively restrict access to state and commercial services, increase monitoring and policing of certain "untrustworthy" people, and even determine what jobs an individual can have. Even more interesting are the institutions that are being created within large companies such as Alibaba that mirror government institutions. Alibaba's disputes between suppliers and customers are resolved by a jury of one's peers and sentences are doled out by impartial judges; these roles are given to users on Alibaba's platform. Judges and juries decide how to split up money held in escrow for disputed transactions made in the system.

This corporate system alleviates much of the strain put on rigid and ineffective legal institutions (*Liu and Weingast*, 2017). Despite being a competitor to the existing bureaucratic system, the government appears to tolerate its existence. It is conceivable and consistent with current trends that these corporate structures might one day inform reform of legal institutions, or be subsumed into the current legal system in China.

The trends outlined above and the relationship between Sina Weibo and government actors in China may represent a potentially transformative shift in how states and corporations interact. More nimble and adaptive corporate structures may help authoritarian governments leverage data to manage and monitor public opinion. Alternatively, corporate profit motives may prove to increase conflicts between state and society, leading to reform or instability.

CHAPTER IV

Reassessing the Targets of China's Online Censorship Apparatus

What are the bounds of the Chinese government's tolerance of online political expression? Following the publication of a series of noteworthy papers by King et al. (2013, 2014), a consensus has emerged contending that the government tolerates political criticism and selectively targets content with collective action potential. Nevertheless, we continue to witness numerous cases of censorship, arrests, and repression of users who post online criticisms, political humor, and discussions of leadership that have little to do with collective action. In this chapter, I demonstrate that the Chinese government has a broader agenda to constrict the space for counter-hegemonic discourse, which includes the suppression of both political criticism and content with collective action potential. By drawing on direct measures of government intent as recorded in leaked censorship documents, I find that although censors frequently target collective action content, they are even more likely to target discussions of leadership and government criticisms.

4.1 Introduction

Since mid-July of 2017, all images and mentions of Winnie-the-Pooh have been scrubbed from the Chinese internet. This is because, according to a popular meme, China's "core leader" Xi Jinping resembles the bear. While censorship of Winnie-the-Pooh in China may seem aberrant, censorship of government-critical content and seemingly innocuous content such as images of tattoos, or pride flags is commonplace.[1] This is puzzling because, according to popular consensus in the political science literature, the government should not be targeting anything but content with "collective action potential" (*King et al.*, 2013, 2014). In this chapter, I demonstrate that the Chinese government's censorship agenda is far broader than the existing literature suggests. While the state cares about content with collective action potential, it cares also about constricting space for anti-regime content, whether or not this content can lead to on-the-ground protest.

Understanding the bounds of the Chinese state's tolerance for political expression is important because crossing these bounds can have real-world consequences for ordinary Chinese citizens. Countless citizens are not only censored, but imprisoned or interrogated for criticisms they write about government leaders (*Tager et al.*, 2017). For example, in April 2017, a man was sentenced to two years in prison for calling the president "steamed buns." By underestimating the importance of criticism, we may be limiting our understanding of broader state repression in China.

Using an original dataset of 8,427 leaked censorship logs from popular social media company Sina Weibo, I compare the distribution of content targeted for censorship to empirical expectations of the collective action potential. I find that, while censors target collective action content at high rates, they target government criticism and discussions of leadership even more.

Finally, I suggest that research on censorship—and state repression more broadly—

[1] See: http://www.webcitation.org/72D1Zfksb

ought to directly measure and not assume government intent. I challenge assumptions that "the state's revealed preferences" can be uncovered by analyzing censorship outcomes and suggest that the randomized experiment in *King et al.* (2014) and the big data analysis in *King et al.* (2013) may underestimate the importance of non-collective action content. Because many non-government actors influence what is and is not censored in China, the outcome of censorship is an inaccurate measure of government intent.

4.2 Collective Action Potential vs. Low Censorship Capacity

The collective action potential hypothesis argues against a "conventional wisdom" that the main target of censorship is criticism of the government. Instead, the authors argue that "the purpose of the censorship program is to reduce the probability of collective action by clipping social ties whenever any collective movements are in evidence or expected" (*King et al.*, 2013, 326). They find that "posts are censored if they are in a topic area with [collective action potential] and not otherwise. Whether or not the posts are in favor of the government, its leaders, and its policies has no measurable effect on the probability of censorship" (*King et al.*, 2013, 339). For an example of collective action potential content, the others cite "posts on a local Wenzhou Web site expressing support for Chen Fei, a environmental activist who supported an environmental lottery to help local environmental protection." This content was not anti-government but was still censored. For government criticism, the authors are less clear about their coding rules. The authors argue repeatedly that the patterns they uncover in the data "seem to clearly expose government intent" (*King et al.*, 2013, 326). Related theories report salutary effects (for the state) of opening space for criticism. Some claim that the government may benefit from watchdogs in media that identify corrupt subordinates and popular grievances (*Dimitrov et al.*, 2013; *Malesky and Schuler*, 2011; *Egorov et al.*, 2009b; *Lorentzen*, 2014). Others claim

that circumscribed spaces for criticism can be used to collect grievances and feedback on governance and policy proposals (*Chen and Xu*, 2016; *Truex*, 2014; *Gueorguiev and Malesky*, 2018).

While the collective action potential hypothesis focuses on a universal government intent behind incomplete censorship, many works focus instead on how and why the state is unable to perfectly control information. Others suggest that the government lacks capacity to censor as completely as it would like due to state-society conflict (*Esarey and Xiao*, 2011; *Diamond*, 2010) or government fragmentation (*Han*, 2018; *Cairns*, 2017). Others suggest that market competition in the media can lead to circumscribed space for critical information to spread (*Stockmann*, 2013; *Miller*, 2018).

4.3 Data and Methods

4.3.1 Leaked Log Data

In this chapter I test the collective action potential hypothesis by analyzing the content of a new database of 8,427 censorship logs from popular social media company Sina Weibo. These logs are a complete set of documents from April 2011 to late 2014 that record censorship orders and daily business at Sina Weibo's content censorship office in Tianjin, China. These documents include government censorship directives, the content to be censored, management decisions to implement or defy these directives, and other general notices to content moderation employees. Logs are meant to share information with content moderation employees working in different shifts in an effort to minimize duplicated management effort. This corpus of documents was leaked to the Committee to Protect Journalists by a former employee working as a content moderator at Sina Weibo.

These data are fundamentally different than the data used in past analyses. Many analyses of censorship have drawn inferences from censorship outcomes by sampling

social media posts and periodically checking to see if that post has been deleted at later stages. In contrast, log data are completely separate from the mechanisms of censorship as experienced on the user end. Censorship logs from Sina Weibo log all government censorship directives they receive. These directives are created by propaganda and public security bureaucracies who proactively decide what is and is not off-limits. Along with the text of directives, Sina Weibo managers write notes to inform employees of how/if they are to implement the directive.

In this chapter, I argue that these data are better equipped to measure government intent than censorship outcomes because they are unadulterated by the many non-government actors who clearly have a say in what is and is not censored on the user end. *King et al.* (2013, 2014) indirectly measure government intent through censorship outcomes (whether or not posts on social media, blogs, and other platforms were censored), claiming that these data expose "revealed preferences through [the government's] censorship behavior," despite the many non-government actors (such as private internet companies) who decide what content is and is not visible. The leaked logs capture government intent before it is distorted by these non-government actors. These data are also well-suited to test the collective action potential hypothesis because they were generated during the same period as data were collected for both papers by King et. al., containing a complete set of logs from 2011-2014, holding the time period of censorship observed by both studies constant.

These logs are from a single social media company, Sina Weibo. As such, there are limits to the external validity of these inferences. That being said, during the time of the analysis, Sina Weibo boasted over half a billion registered users and was ranked in the top 3 social networks by monthly active users. Many of the logs indicate as well that the same directives sent to Sina Weibo were also sent to competitor Tencent. Together, Sina and Tencent represent the vast majority of the social media market. Directives to Sina Weibo come from Beijing-based regulators. The largest number of

ICPs are registered in Beijing so any bias toward Beijing in this analysis does not largely impact generalizability of inferences. This is also less of a concern since the topic distributions uncovered in this analysis are similar to those measured by *Cribben et al.* (2018) from "internal documents from Hunan Province" during the Hu Jintao period (2009-2010).

4.3.2 Coding Procedure

Along with two research assistants, I manually labeled each log according to the content of posts targeted for censorship. In total, I kept track of 17 top-level topic categories and 51 more specific secondary categories. Nearly all topic categories have high intercoder reliability (see Table A.1 in the appendix). Topic categories were developed inductively, starting with categories existing literature posited were of theoretical interest. New categories were added when they could not neatly fit into any of the existing categories. These topic categories include mentions of government leadership, government criticism, collective action, and corruption, topics of theoretical interest in the literature reviewed above. A brief description of content categories relevant to this analysis can be found in Table 4.1. Complete coding diagrams can be found in Section A.1.2 in the appendix.

Labeling censorship logs took nearly three years of continuous work due to the lack of structure in the raw data. Log data included a mix of data formats (images, text, video), and were written in a jargon-heavy shorthand that is difficult to understand without training. All of these hurdles made it infeasible to use automated methods of text analysis. Raw documents from the leak had no clear and consistent delimiters between logs, so approximately 10,000 logs needed to be manually segmented before further processing. After this, approximately 2000 duplicates were identified using the Smith-Waterman edit distance algorithm (*Smith and Waterman*, 1981)and reviewed manually.

I hired two research assistants to read and categorize each log. All coding rules were provided to research assistants in flow-chart form. Research assistants were instructed to follow these flow charts as they coded. In total, research assistants and I made 573,036 individual content categorizations. Before coding began, I defined and diagramed all content categories, trained each coder, and performed periodic intercoder reliability checks. In order to achieve adequate intercoder reliability, each coder needed months of training. The two final coders were selected from an initial pool of 6 coders who went through the complete training process and labeled at least 1000 logs on their own. These coders were selected based on their demonstrated understanding of the coding scheme and their consistent quality of work.

I made a concerted effort to conform to the coding rules for concepts defined by *King et al.* (2013), but the original coding scheme required some adjustments due to the vagueness of concept definitions and differences in the nature of categories (this analysis uses mixed membership). The concept "collective action potential" discussed in *King et al.* (2013) is very broad, encompassing "any event that has the potential to cause collective action." In my own coding scheme, I worried about the meaningfulness of such a category since almost any event or individual has some potential to cause collective action. It was not clear whether the authors drew a line for concept membership above a certain propensity to cause collective action. Nonetheless, the authors define "collective action potential" as events that belong to one or more of the following sub-categories: protest, individuals/activists, and nationalism.[2] I measured each of these subcategories of "collective action potential"—as defined by *King et al.* (2013)—separately. Coding rules for each of these categories can be found

[2]The full definition is as follows: "events which (a) involve protest or organized crowd formation outside the Internet; (b) relate to individuals who have organized or incited collective action on the ground in the past; or (c) relate to nationalism or nationalist sentiment that have incited protest or collective action in the past." I measured each component of "collective action potential" as defined by *King et al.* (2013) (a) as "protest," and (b) as "social activism," both under an umbrella category "collective action." So as not to miss any other forms of collective action, I also measured concepts "strikes," "petitions," and "social groups." I measured (c) separately as "nationalism."

in Section A.1.2 the appendix. To be conservative, I approximated collective action potential using a broader category that encompassed each of these components and a few additional sub-categories. For clarity sake, I call this concept simply "collective action." This category includes content that either "1) mentions or implies an event where a group of people took action together to achieve a common objective, or 2) mentions or implies an individual or group of individuals who are advocating on behalf of a social, religious, or ethnic group." This is a broader conceptualization of collective action than *King et al.* (2013). It includes online collective actions such as using candle emojis to participate in a digital vigil after the death of Nobel Peace Prize Winner and dissident Liu Xiaobo, or the collective signing of online petitions and political documents such as Charter 08 for which Liu was sentenced to an 11-year prison term. As such, if anything, the differences in coding scheme will over-estimate the prevalence of "collective action potential content" as originally defined by the authors.

The coding rules for the concept "government criticism" were not provided in the text or supplementary materials. Though I planned on reverse-engineering coding rules from post-level data to address this problem, unfortunately post-level replication data are unavailable due to storage issues.[3] Instead, I defined government criticism as content that satisfies membership requirements for the "government" parent category and either "(a) speaks ill of, criticizes, or ridicules government leaders or their families, government policies, or government institutions or (b) includes instructions that mention "negative" content or content that 'attacks or 'mocks"' (see Table 4.1). Other concepts measured in this content analysis are diagrammed and described in detail in Section A.1.2 in the appendix.

[3] *Gueorguiev and Malesky* (2018) were unable to obtain replication data and instead performed a "pseudo-replication" on the aggregate data rather than the raw data. They find that many of the criticisms that make it into the sample are solicited by the Chinese government in the first place in processes known as "public opinion consultation" and "public opinion supervision."

Table 4.1: Brief Description of Topic Categories

Category	Parent Category	Description
Government	-	The content mentions or implies a Chinese government institution, organization, or bureaucracy, a Chinese government official of any rank or position, their family members or their partners/mistresses, a Chinese government policy, or a Chinese state-owned enterprise (SOE).
Government Leadership	Government	The content mentions or implies a Chinese government official of any rank or position, their family members, or their partners/mistresses.
Government Criticism	Government	The content speaks ill of, criticizes, attacks, mocks, or ridicules government leaders or their families, government policies, or government institutions.
Political Humor	Government	The content involves mockery, humor, or satire of government leaders or their families, government policies, or government institutions.
Corruption	Crime	The content mentions or implies any of the following: 1) misuse of local government office or local government funds, 2) sexual misconduct of local government officials 3) a local government official and/or his/her family financially benefiting from a government post.
Collective Action	-	The content either 1) mentions or implies an event where a group of people took action together to achieve a common objective, or 2) mentions or implies an individual or group of individuals who are advocating on behalf of a social, religious, or ethnic group.
Social Activism	Collective Action	The content mentions or implies advocacy on behalf of a social group. Content can either mention a group directly, or mention a member of a social group (i.e. Chen Guangcheng).
Protest	Collective Action	The content mentions or implies a street protest, march, or collective walk.

4.3.3 Empirical Expectations

If the collective action hypothesis is correct, we should expect the the vast majority of censored content to be related to "collective action" with little-to-no censorship of posts related to discussions of crime, government criticism, or discussion of leaders. In contrast, a large number of logs cases related to discussions of crime, government criticism, or discussion of leaders would provide evidence against a strong version of the collective action hypothesis.

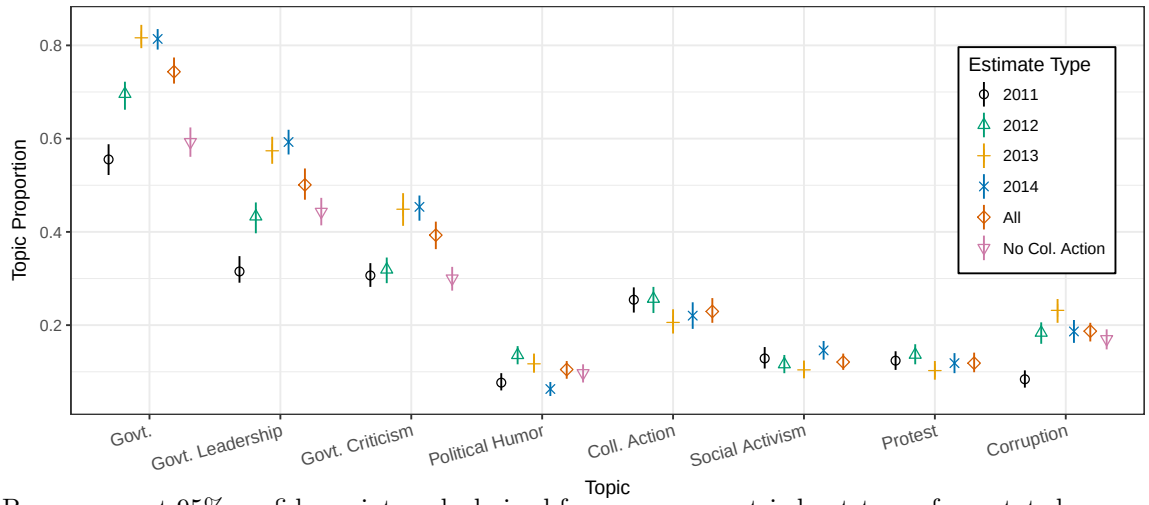
4.4 Results

Contrary to empirical expectations of the collective action potential hypothesis, I find that discussions of leadership are targeted most frequently, that government criticism is almost twice as likely to be targeted for censorship as is collective action content, and that corruption, collective action, crime, and a wide range of other topics are censored at similar rates. Further, it appears that government criticism and discussions of leadership became more frequently targeted in the Xi Jinping era (2013-) while collective action content became less frequently targeted, a break from the Hu Jintao years covered in the data (2011-2012). The distribution of content targeted by governments in the log data provides clear evidence that the collective action potential hypothesis overstates the importance of collective action potential and understates the importance of government criticism, discussions of leadership, discussions of corruption, and discussions of crime.

Government-related posts are targeted most frequently in logs, meaning that government directives to Sina Weibo are most frequently about the government (74.30% of logs). Government directives target collective action content in only 23.06% of logs. Discussions of government leadership are the second largest topic, representing 50.01% of logs. Government criticism is targeted much more frequently than collective action, at 39.29%. Political humor makes up 10.41% of logs. Posts about corruption are nearly as common as collective action posts, and make up 18.68% of logs. The distribution of government, collective action, and corruption topics are visualized in Figure 4.1; the distribution of all topics representing greater than 5% of logs can be seen in 4.2. Below I perform a few robustness checks to make sure that the distribution of log content is not driven by mixed membership of categories, unusual events happening in particular years, or due to the influence of potential over-censorship on the Sina Weibo platform.

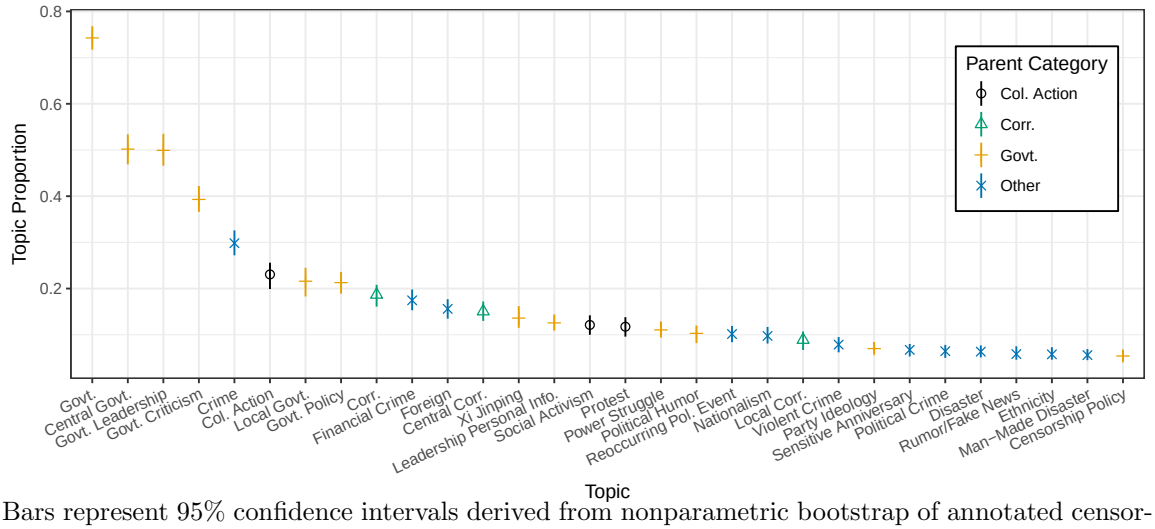
Each log can belong to several categories (mixed membership). It is therefore

Figure 4.1: Topic proportions overall, by year, and without mixed CA membership



Bars represent 95% confidence intervals derived from nonparametric bootstrap of annotated censorship logs.

Figure 4.2: Topic proportions for category proportions over .05



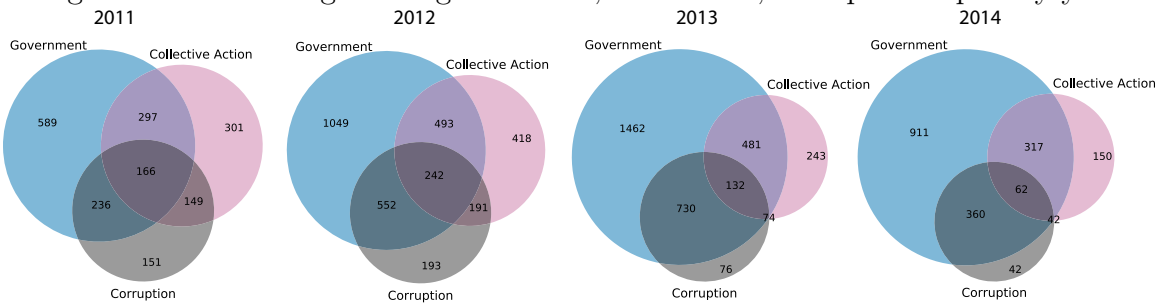
Bars represent 95% confidence intervals derived from nonparametric bootstrap of annotated censorship logs.

possible that the high prevalence of censorships related to government content was driven by a mixed membership with collective action. I therefore recalculated the frequencies of each category, excluding logs that have mixed membership with the collective action category. After adjusting proportions for mixed membership, results do not appear to be driven by mixed membership (see Figure 4.1). This highlights an important difference between both of King, Pan, and Roberts' approach to topic

categories and the one used in this analysis. In their work, collective action and government criticism are treated as mutually exclusive categories. In this analysis, 9.41% of logs target posts that contain collective action content *and* government critical content. Since the authors do not report any coding rules for the government criticism category, it's hard to know how or if a line was drawn between the two concepts.

Finally, because in late 2011 and 2012, an unusual number of high-profile government-related events took place, I examined logs across years to test that these findings are not driven by the unique events of late 2011 and 2012, namely the Bo Xilai affair, the 18th Party Congress, and speculation about leadership transitions as Hu Jintao retired from office. This might drive a lot of criticism or discussions of leadership in the leaked log data. In Figure 4.3 the results by year show the opposite trend: collective action and corruption were more heavily targeted in directives in 2011 and 2012 than they were in 2013 and 2014. In later years under President Xi Jinping, it appears that government leadership and government criticism became much more frequently targeted.

Figure 4.3: Venn diagram of government, col. action, corruption topics by year



4.5 Discussion

In this chapter, I tested the collective action hypothesis by measuring the topic proportions of content targeted for censorship from a leaked dataset of censorship logs

from Sina Weibo. I measured the distribution of topic categories by manually labeling content according to several topic categories of theoretical interest. I found that contrary to the empirical expectations of the collective action potential, discussions of crime, government criticism, and discussion of leaders are more frequently targeted for censorship than collective action content. In addition, a diverse array non-collective-action topics are censored at non-negligible rates.

4.5.1 Beyond Collective Action Potential

If collective action potential is not exclusively targeted, then what explains the diversity in censored content? The literature offers two main explanations for this diverse array of content targeted for censorship. First, this diversity can be explained by the state's preference for reducing the prevalence of counter-hegemonic discourse by targeting influential people rather than categories of content. Second, this diversity can be explained by diversity in actors involved in censorship who often have conflicting preferences.

Past work suggests that the Chinese Communist Party controls information and prevents challenges to it's monopoly on power by reducing the influence of—or subsuming—organizations and ideologies that are counter-hegemonic (*Schurmann*, 1966). We might then expect that the state would target opinion leaders and groups of individuals who are engaging in counter-hegemonic discourse, even in cases where this discourse has no collective action potential. *Gallagher and Miller* (2018) find that users with high follower counts and retweets are more likely to be reported to the authorities by social media companies regardless of the topic of relevant content shared. This explanation extends the collective action hypothesis by re-situating it within an earlier hypothesis by *Schurmann* (1966) about how the Chinese Communist Party controls society through controlling ideology (targeting counterhegemonic discourse, i.e. criticism of government, leadership, official party line) and organization (target-

ing collective action potential by embedding the party in all organizations and social groups).

Other work cites the many levels of government, bureaucracies and individuals with a stake in censorship (*Miller*, 2018). The diversity of preferences of these government actors may explain the diversity in content targeted for censorship. *Cribben et al.* (2018) find that central and local preferences for censorship differ substantially. Diversity in actors involved in censorship was acknowledged in *King et al.* (2013) as a potential limitation to claims of government intent.[4] In traditional media, the diversity of censorship objectives in China has been well-studied (*Stockmann*, 2013; *Brady*, 2009; *Shambaugh*, 2007)

4.5.2 Why Government Intent Should Not be Inferred from Censorship Outcomes

In *King et al.* (2013, 2014), the collective action hypothesis was tested by measuring censorship outcomes and comparing the rate of censorship of collective action to the rate of censorship of government criticism. These quantities, however, may not accurately measure the intent of the government. For censorship outcomes to accurately reflect government intention, one must assume that: 1) the government acts uniformly as the sole actor involved in censorship (no non-government actors are involved); 2) the government has a single, coherent censorship strategy; and 3) the government can effectively delegate to agents of censorship. Below, I outline a series of specific reasons why these assumptions are invalid, and why it is unlikely that the research designs of *King et al.* (2013, 2014) accurately measure concepts of theoretical interest.

Governments are not the only decision-makers behind what is and is not censored

[4]The authors concede that "in those instances when different agencies, leaders, or levels of government work at cross purposes, even the concept of a unitary intent or motivation may be difficult to define, much less measure."

in China. In China, the government delegates to private internet companies to censor on their behalf, and these internet companies have a lot of power. They routinely disobey directives in pursuit of profits (*Miller*, 2018). Additionally, these companies censor for reasons unrelated to government concerns. They censor harassment, pornography, and spam, as social media companies routinely do even outside of China. Evidence from log data and interviews with journalists and researchers show that social media companies, and sometimes rogue employees, frequently censor on behalf of celebrities or friends (*Wang*, 2016a,b; *Cairns*, 2017).[5] Private internet companies often censor content on their platform that is detrimental to the company's interests, such as unfavorable reports about the company's IPO in the United States (*Miller*, 2018). While *King et al.* (2014) acknowledge that corporate actors are involved in the process, they do not consider whether their involvement in the process could result in outcomes deviating from government intent; instead, they treat corporate delegation as a black box.

Ordinary users also play a large role in determining what is censored. From censorship outcomes alone, it is impossible to determine whether a post was deleted as an act of censorship or whether the original poster deleted it. As such, measures of "censorship" outcomes could include instances of self-censorship. Some individuals may be pressured by family, friends, or coworkers to delete content or they may have second thoughts after posting about the social desirability of the content they just shared. In these cases, self-censorship can often be conflated with government-directed censorship when looking at censorship outcomes. Content at internet companies in China is flagged for manual review by ordinary users who click "report" buttons. These crowdsourcing systems identify objectionable content not from government directives, but from signals of what netizens find socially undesirable.

[5]The source of Sina Weibo's leak admitted that while at Sina, they "helped friends and strangers get back accounts that had been removed and told them how to walk around sensitive words. I also helped [influential users] find out which government agencies ordered the removal of their accounts" (*Wang*, 2016b).

The above ways in which censorship outcomes give an inaccurate read on the intent behind censorship shows the significance of understanding state repression starting at decisions and reading intent forward. This observation has been made previously by *Capoccia and Ziblatt* (2010), who claim that comparative historical analyses of democratization can overlook important determinants of democratization by reading backward rather than reading forward. Backward inducing intent from outcomes of censorship overlooks non-government actors and internal government dynamics that affect what is and is not censored, often in ways that circumvent initial intent.

4.6 Conclusion

In this chapter, I demonstrated that the Chinese state's tolerance for political expression tolerates is much narrower than is currently appreciated. The state's broader agenda includes the suppression of both political criticism *and* content with collective action potential. This corrective of *King et al.* (2013, 2014) is significant because it shows that one or both of the following is true: 1) the Party-state is weaker than we assume and lacks capacity to target censorship, as has been previously assumed; 2) the Party-state cares a great deal about counter-hegemonic discourse, and is focused primarily on reasserting party control over public spaces it does not currently control.

Methodologically, this chapter stresses the significance of understanding state repression starting at government decisions and reading intent forward. By focusing on outcomes of censorship and reading backward, *King et al.* (2013, 2014) failed to account for many ways in which censorship outcomes might not reflect government intent, as private actors often countervail government interests, and ordinary users participate in censorship. While discussions of the government logic behind censorship are important, backward inducing intent from the roles the system seems to play should be avoided, and can lead quickly to functionalism.

CHAPTER V

Automated Detection of Chinese Government Astroturfers Using Network and Social Metadata

5.1 Introduction

Astroturfing is the promotion of an opinion or propagation of information through fabricated "grassroots" behaviors and/or "social movements." It is a tactic that has been used by tobacco and oil companies to promote support for policies that are advantageous to their interests.[1] Authoritarian governments and political parties in democracies make use of similar tactics to suppress discussion or guide opinion. Government astroturfers in China, popularly known as the "Fifty Cent Party" (五毛党)[2], are employees of a wide range of government bureaucracies who are tasked with "guiding opinion" online. They post pro-regime messages on social media platforms, deliberately hiding their identity, with the goal of appearing to be ordinary citizens.

While there is a growing literature in political science on bot detection (*Stukal et al.*, 2017), very little work has been done to detect astroturfers. While these bot-detection methods must discriminate between humans and machines, astroturfer

[1]Evidence of campaigns and their effectiveness can be found in work by *Cho et al.* (2011).

[2]Government astroturfers in China are colloquially referred to as members of the "Fifty Cent Party" (五毛党). They are so called because they are purportedly paid 0.5 RMB per post (0.07 USD) to post pro-regime commentary on social networks, online news channels, and other websites with user-generated content. Their official title is usually some variant of "internet commentator" (网络评论员)

detection must differentiate one class of humans from another. Astroturfers, unlike bots, are humans, usually posting manually, and often crafting custom messages. Because the practice of government astroturfing involves hiding one's identity and appearing to be an ordinary user, differentiating an astroturfer from an ordinary user is a difficult task. There has been very little scholarship dedicated to government astroturfing, and there have been no successful attempts to identify government astroturfing in China. Computer scientists have tried to use unsupervised methods to detect activity from government astroturfers, but they relied on text data and made many assumptions about the text content of astroturfer posts. These methods were unsuccessful, and researchers "located no evidence that any of [observed] users are [Fifty Cent Party members]" (*Yang et al.*, 2015).[3]

A key obstacle to the study of political astroturfing is the difficulty of data collection. So as to prevent researcher-induced bias, analysis of astroturfing often requires empirical ground truth[4] data. Because of limited available data, researchers have sometimes relied on a "you'll know it when you see it" approach to identifying astroturfers. For example, ethnographic research by *Han* (2015b,a) identifies astroturfers by searching for language that "[smells] strongly of official propaganda" (*Han*, 2015b). Though these users were likely to have indeed been astroturfers, there is a small chance that some of those users' opinions were genuinely in line with official propaganda and they were commenting independently, and not on behalf of a government organization. In China these users are called the "Volunteer Fifty Cent Party" (自干五). Though Han's analysis was careful, this "you know it when you see it" approach to astroturfing may introduce less careful researchers' biases or preconceptions about what astroturfing looks like. These biases may reflect popular conceptions of

[3]This is likely because text content alone can not adequately discriminate between government astroturfers and ordinary citizens in the same way that non-text features can.

[4]"Ground truth" refers to information that has been gathered empirically rather than through inference. In this analysis, "ground truth" refers to data where the identity of the commentator—astroturfer or not—can be observed. If a model's predictions resemble empirically observed "ground truth," this suggests that inferences from a model are of good quality.

astroturfing as depicted in the media and academic literature. Many assumptions about the Fifty Cent Party appear to be myths, most notably, the assumption that astroturfers are paid piecemeal (50 cents per post).

Because of the difficulty of differentiating users with genuinely pro-regime opinions from astroturfers, recent works have advocated the use of ground-truth data, where the identity of government astroturfers has been somehow uncovered. That is, leaks or public disclosures have identified that the true source of comments is astroturfers and not ordinary netizens. *Keller et al.* (2017) make use of publicly disclosed astroturfing data from the Park campaign in South Korea to describe the behavior of astroturfers during and after a highly contested election. In another ground-truth-based study, *King et al.* (2016) estimate the number of astroturfer comments made each year in China. They extrapolate from astroturfer comments reported to managers via email as found in a cache of hacked emails from a local district Propaganda Department in China. Research by *King et al.* (2016) asserts that astroturfing is not about persuasion, and more about distraction through "cheerleading."

The purpose of this paper is simply to outline a method for detecting astroturfers. However, work in progress that analyzes the content of posts identified using these methods does not support the assertions of *King et al.* (2016). Instead, these preliminary findings suggest that the purpose of astroturfing is to respond to "public opinion emergencies" through agenda-setting, and dilution of negative sentiment through posts with "positive energy (正能量)." More work however needs to be done to fully understand whether astroturfing is effective and what the state's objective behind astroturfing is.

Though studies that make use of ground truth data avoid the problem of researcher-induced biases and have no trouble disambiguating ordinary users with pro-government positions from astroturfers, the data used for these analyses can come with their own biases. Exclusively relying on these ground-truth datasets may limit our analysis to

the moment in time captured by leaks or disclosures. Without a method of detecting astroturfers, researchers risk becoming dependent on rare leaks or public disclosures to study astroturfing. This forces researchers to either study astroturfing using cross-sectional analyses, or assume that inferences from one cross-section can be applied to others. This becomes a problem when leaked or publicly disclosed data are regionally biased, involve highly specific subject areas, or cover a short span of time.

In this chapter, I outline a method of identifying government astroturfers that can be validated usign ground truth data, but does not rely on it exclusively for the analysis. This method leverages metadata that is commonly provided alongside text features in social text scraped from the internet. Because behavioral patterns are encoded in metadata, researchers can draw upon documentary sources to create rules that differentiate astroturfer behavior from the behavior of ordinary users. To address possible researcher induced bias, researchers can use ground truth data to validate rules for discriminating ordinary users from astroturfer users. This approach involves 1) identifying work and behavior patterns that differentiate astroturfers from ordinary users, 2) retrieving likely and unlikely astroturfer texts from large text corpora using these rules, 3) training a binary text classifier with likely and unlikely astroturfer texts, and 4) validating this classifier by comparing model predicted outcomes to ground truth outcomes.

I utilize this approach to detect astroturfers in a large database of 70 million news media comments from 19 popular news outlets that vary in their state-affiliation, level of commercialization, and region. First, I identify behavioral patterns of astroturfers using a in-depth study of a corpus of government documents and training manuals (*Miller*, 2016). I create several rules that discriminate between astroturfers and ordinary users using these sources and analyses. Second, I retrieve comments that satisfy the rules outlined in the first step. Third, I train several binary classifiers to discriminate between the comments retrieved in the second step from their complement in

the corpus. Finally, I use ground truth data of leaked astroturfer comments from the Zhanggong Propaganda Department to validate these models. Each classifier predicts astroturfer comments from the Zhanggong leak with greater than 90% accuracy.

5.2 Identifying Government Astroturfers

In order to avoid researcher-induced bias, I use metadata rather than text content to infer whether a user is an astroturfer or an ordinary user. In my data, metadata includes IP address, post time, social network data from Weibo (a Chinese social media service similar to Twitter), comment likes, usernames, user locations, etc. These metadata can be used to analyze comments without any assumptions about the syntactical or dictional content of the comment text. Instead, using what is known about government astroturfers' network behavior and modal job responsibilities, researchers can look for empirical patterns one would expect from only astroturfers and not ordinary users.

If researchers can identify patterns in metadata that can convincingly discriminate between ordinary and astroturfer users, they can make inferences about the identity of astroturfers without relying on "ground truth" data. Inferences however, will only be certain in the handful of cases where astroturfers identify themselves in their social media handles as Figure 5.3 shows some doing. This is why, instead of starting with ground truth data, I use ground truth data to validate my metadata-based detection approach.

5.2.1 What We Know About Government Astroturfers

Documents about tactics, job responsibilities and institutional structures can be readily found on the websites of various bureaucracies, and party instruction manuals and textbooks are widely available for purchase in Chinese bookstores. These documentary sources, first used in *Miller* (2016) provide useful information about the

process of government astroturfing in China.

Though bureaucracies and government organs such as the Central Propaganda Department, the Cyberspace Administration of China, and its parent organization, the Central Leading Small Group for Internet Security and Informatization are notoriously secretive, the process of government commentating is actually not very sensitive. Government commentating is routinely and openly discussed at all levels of government and in the press (*Zhang*, 2011; *Global Times Editorial Team*, 2016), and has even been addressed, though indirectly, by President Xi Jinping (*Huang and Zhai*, 2013; *Xinhua*, 2016). Public acknowledgement, such as the Henan Public Security Bureau's press release boasting about hiring 100 government astroturfers, is quite common.[5] Moreover, it is not uncommon for Chinese netizens to support and condone this practice. The Chinese Government frames this practice as a means to combat hostile Western forces, protect Internet sovereignty, and guide public opinion.

Commentating teams exist throughout the vast web of Chinese bureaucracies.[6] Core commentating teams, when needed, can draw on individuals serving under other bureaucratic functions to aid in an urgent or more intensive campaign. Astroturfers usually are a part of public opinion monitoring divisions, and exist alongside structures responsible for surveillance. When public opinion analysts uncover a potential threat, commentating teams are directed to implement detailed contingency plans, responding with a carefully crafted and unified message to the specific public opinion event (*Zou and Su*, 2015).

After a close reading of government documents, manuals, and textbooks for internet commentators and public opinion monitors, I outline several common behavioral patterns of astroturfers that one would not expect to see from ordinary users. To ensure that the patterns described in manuals reflect what we see in practice, I compare

[5]See the press release (in Chinese): http://goo.gl/rgz4xn

[6]Agricultural bureaucracies, tourism bureaus, propaganda departments, information offices, police departments, public security bureaus, prisons, and Communist Youth League organizations all have their own commentating teams, usually in core teams of two or more individuals.

the content of manuals to a leaked email archive from the Zhanggong Propaganda Department and confirm that these patterns are a good measure of astruturfer behavior in practice. These leaked emails include spreadsheets, screenshots, word documents, and text files of astroturfer comments that are reported to managers during district astroturfer campaigns. These data were very messy and required manual and automated methods to process and clean. These data are quite biased, and are likely not representative of how astroturfing works nationally. Zhanggong is not representative of China. It is a district in the small, prefecture-level city of Ganzhou, a relatively poor city. This ground truth dataset contains approximately 40,000 astroturfer comments from 8 distinct astroturfing campaigns, many of which are local PR campaigns that are not nationally salient. The media sources targeted by commentators are often local or esoteric. Though these data are biased, and can only tell us so much about the process of astroturfing ouside of Zhanggong, it allows for empirical validation of government astroturfer behavior identified in documentary sources and later used to build astroturfer detection models.

These text sources suggest that there are no specific bureaucracies tasked with commentating. Instead, I find examples of commentating teams in almost every conceivable bureaucracy. These documents also seem to contradict common understanding that this practice is done in people's spare time, and that they are paid piecemeal. Instead, the documents suggest that nearly all astroturfers are office workers. Additionally, though the practice is not uniform across bureaucracies, certain regularities appear in more developed bureaucracies. For example, astroturfers are often aided by sophisticated monitoring software developed by private companies (Goonie) or state media outlets (People's Daily's Media Opinion Monitoring Office). I have identified over 100 such opinion management services operating in mainland China. Bo Mai confirms this trend in his work on sub-national governments' budgets for surveillance technology in China (*Mai*, 2016). Astroturfers are often full-time workers with spe-

cific titles such as "news spokesperson (新闻发言人)," "internet commentator, (网络评论员)" "public opinion analyst (舆论分析员)," etc. Though these job titles are common, it does not mean that every government astroturfer works full time. Some are called upon only during crises, and when they are not needed, they return to their primary responsibilities. During crises, organizations such as the Communist Youth League encourage members to volunteer as astroturfers. For example, the Guizhou Province's Communist Youth League's guidelines for astroturfer teams stresses the need to "establish a quick mobilization system: Be able to mobilize 20% of members in 3 hours, 50% of members in 24 hours and 80% of members in 72 hours" (*Anonymous*, 2014).

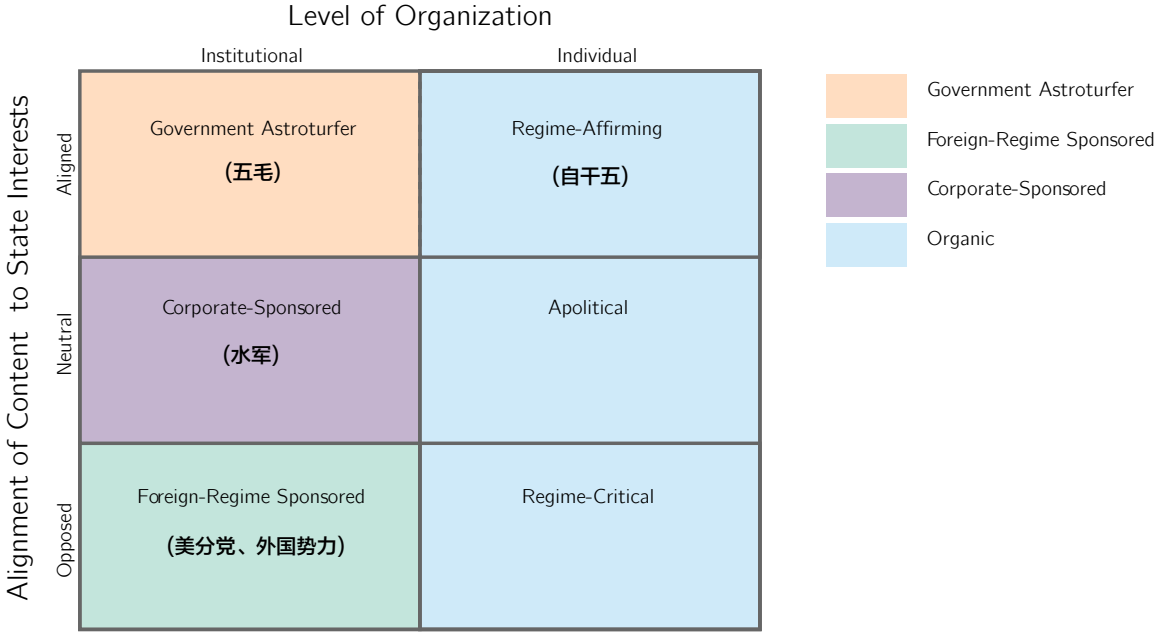
Because Propaganda Department leaks from a single level of government are not generalizable to the practice of astroturfing throughout China's vast bureaucracy, I refrain from using data from available leaks (except for final validation of classification models), and build a model based upon common astroturfer behaviors that are recorded in comment metadata. The behavioral patterns I use are parsimonious and consistent with nearly all official descriptions of government commentating work as reflected in aforementioned documents.

5.2.2 Labeling Observations

I label comments (government or non-government) using common behavioral patterns and job responsibilities of government astroturfers, as inferred from documents and records available on the practice throughout China. In crafting these rules, the goal was to use modal behaviors of government astroturfers so as to capture the many diverse practices of commentating across different bureaucracies and government ranks. When combined together, these rules make it extremely unlikely that a non-government comment would be misclassified as a government comment.

5.2.3 What are the Defining Attributes of Government Astroturfer Behavior?

Figure 5.1: Types of Online Commentary



Based on how the process of commentating is described in government documents, I create a typology of online commentary to describe how I separate real user commentary from government astroturfer commentary (see Figure 5.1). I start with two dimensions of online speech that are most discriminating between the behaviors of ordinary internet users and government astroturfers: 1) the level of organization of posting behavior, and 2) the alignment of content to state interests. Using this typology, we can define astroturfing commentary as content that is aligned to state interests and is produced in a systematic, institutionalized way. A more detailed explanation of this typology can be found in the appendix.

5.2.4 Rules for Automatically Labeling Comments

With the key dimensions that discriminate between normal and astroturfer commentary in mind, I outline the rules I use to search for government astroturfer com-

ments. For each rule, I discuss 1) the *typological classes* into which the rule separates comments, 2) *documentary evidence* from central government manuals and local government documents that confirm behaviors or working procedures targeted by the rule, and 3) *empirical evidence* of the rule's discriminating potential using leaked government astroturfer posts from the Zhanggong Propaganda Department to measure government astroturfer behavior, and both my database of commentary and a random sample of Weibo posts as a baseline for normal user posts. The procedure below will automatically label posts for a training set that will be used to predict comments in the database that lack complete metadata. These rules will select a subset of all comments belonging to government astroturfers that will be used to train a classifier. This classifier will provide an estimate of the proportion of government astroturfers in news comment sections online, the topic distribution of astroturfer-targeted content, and the bureaucratic affiliation of government astroturfers.

1. **There are two or more users posting from the comment's IP address**

   Typological Classes: As described in the previous section, government astroturfers work in office environments at a wide variety of bureaucracies. Employees using the same router, as is common in office situations, almost always share the same IP address. Government astroturfers, because they are more "institutional" than "individual" should then be observed at IP addresses with posts made by several users.[7] *Documentary Evidence:* In the corpus of public and leaked government documents, nearly all documents that included employee rosters either listed several employees with the title "internet commentator" or listed several employees whose responsibilities include monitoring and guiding opinion. It is clear based on evidence from the leaked email archive and these documents that government commentating is almost always done in team

[7]This will be true unless employees share a single account. This would defeat the purpose of comment astroturfing because the entire point of astroturfing is to create the appearance of mass support or unity in opinion.

settings. Moreover, official party cadre manuals, textbooks, bureaucratic job descriptions, and employee responsibility documents indicate that bureaucrats working full-time in other functions are mobilized as needed to comment alongside the core full-time public opinion monitoring and commentating employees. For particularly urgent "public opinion emergencies," this trend will be even more pronounced. This practice is described in detail in several manuals and textbooks on Internet commentating (*Zou and Su*, 2015; *Gao and Zhang*, 2011). *Empirical Evidence:* The emails from the Zhanggong Propaganda Department leak confirm this behavior. There are 233 Weibo accounts listed in URL form in the leak, and there are several news stories and events identified in the leak that astroturfers targeted simultaneously from different accounts. Of the commentating reports to management that include user information, nearly all of them include posts made by several users at the same work unit (单位). Among these documents, there is an average of 11 separate users working on the same commentating task. Additionally, there is ample evidence that individual astroturfers make use of several accounts, posting from them at the same time to increase the appearance of "grassroots" support (*Ai*, 2012). In the Zhanggong leak, several documents confirm this trend, with some users utilizing as many as 50 different accounts at the same time.

2. **There is an unusual volume of posts attributed to the IP address (more than 20 posts)**

   Typological Classes: This rule helps discriminate between "institutional" and "individual" accounts, and is particularly helpful at screening out public WiFi networks that may have a high number of usernames associated with an IP address, but do not have a large volume of posts that would be characteristic of a large commentating campaign. *Documentary Evidence:* In rosters of employees, most government offices with astroturfer teams usually have at least

2 employees working full time on public opinion management. Because of this trend, for government IP addresses, there are likely more posts per IP address than an ordinary internet user who chooses to participate in article discussion threads. Moreover, government astroturfers are often encouraged and evaluated based on the number of posts they make (*CCP*, 2013), often receiving a certain number of points per post. *Empirical Evidence:* In the Zhanggong Propaganda Department leak, email reports to management detailing commentating work have an average of approximately 30 posts. In the leak, there is also evidence of users posting several thousand comments during a "public opinion emergency." In contrast, the average number of posts per IP address in my entire dataset, which can be seen as a baseline for normal commenting behavior is 1.85. I set the threshold at 20 posts, an approximate lower bound for the number of comments reported for campaigns in the Zhanggong leak.

3. **Posts belonging to the IP address are sentence-length (post length, in characters are on average > 20).**

   Typological Classes: This rule helps discriminate between "institutional" and "individual" accounts, capturing the more conscientious behavior of commentators and filtering out more ad-hoc, unvarnished posts of ordinary commentators. *Documentary Evidence:* In government documents a great deal of attention is paid to the quality of posts made by government astroturfers. Documents that outline government astroturfer job responsibilities usually require ideological purity and sophistication of post content. Additionally, astroturfers are consistently required to be well-informed on relevant issues. All messages are to conform to the instructions given by managers, and are reported back in emails. These requirements make it less likely that a post coming from a government astroturfer will be short. Government astroturfers in many circumstances are required to make posts that are longer than a certain number of characters

(*CCP*, 2013). *Empirical Evidence:* In the leaked Zhanggong archive, the average post length was around 30 characters, and most are complete sentences. To ensure most comments from a user are written in complete sentences, I set the threshold to an average of 20 characters.

4. **Weibo accounts associated with this IP address follow or are followed by at least one government account.**

   Typological Classes: This class separates commentary that is "aligned" with government interests from commentary that is "neutral" or "opposed" to government interests. *Documentary Evidence:* Government offices in my database of government documents frequently seek to gain large numbers of followers, as this is an important metric of local government performance (*Peoples Daily Weibo Data Center*, 2015). Moreover, managers also may follow individuals so as to supervise their work. *Empirical Evidence:* Because astroturfers work in office environments, management often requests that employees follow the Weibo account of their work unit or other relevant office(s). This is evident in the Zhanggong Propaganda Department leak, which includes several spreadsheets of employees accounting for which of them have followed government accounts as required. Because there are over 200,000 Weibo accounts to classify, I built a classifier to automatically determine whether Weibo accounts are government or non-government accounts using a subset of labeled data. The practice of government astroturfers following and being followed by government accounts is empirically observable in the leaked Zhanggong Propaganda department emails. According to the decisions of the government Weibo classifier (described in detail below) of the 233 accounts belonging to government astroturfers in the leaked Zhanggong Propaganda Department dataset 127 (54%) of them follow, or are followed by government Weibo accounts. As a baseline for normal behavior, I drew a random sample of Weibo accounts (*Fu and Chau*,

2013) and found that in contrast, only 10% of them follow or are followed by government accounts.

***

If all of these rules are jointly satisfied, it is highly unlikely that government documents returned by the search would not come from government sources. Predictive accuracy of models trained on this subset of comments (94.1% accuracy in predicting Zhanggong Propaganda Department comments) adds supportive evidence that each of these rules is discriminating between ordinary comments and comments from government astroturfers. The resulting set of predicted government comments nearly all appear to have a clear pro-government agenda.

5.3 Government Weibo Account Classifier

Figure 5.2: Profile Pictures from a Random Sample of Predicted Government Weibo Accounts

The most difficult part of building the training set mentioned in the previous section was identifying government Weibo accounts in a astroturfer's social networks

(rule 4). A sizable subset of comments in the dataset (roughly 4 million) contained links to the authoring user's Weibo account. This is because in order to comment on Sina News, one must sign in with their Sina account, which is also used for the Weibo microblogging platform. In order to determine whether rule 4 of the labeling process is satisfied (that at least one of the IP address's users follows or is followed by a government account), I needed to determine if each Weibo account in each user's social network is a government account, i.e. an account that is the official Weibo account of a government bureaucracy or government organ, or an account of an idividual who identifies as a manager or leader within a bureaucracy or government organ. State-owned enterprises are excluded. To find these accounts, I construct a support vector machine (SVM) classifier[8] trained on manually labeled Weibo accounts with class labels:

$$\mathcal{C}(x_i) = \begin{cases} 1 & x_i \text{ is a government account,} \\ 0 & \text{otherwise.} \end{cases}$$

5.3.1 Features

The following features are extracted from Weibo account data and are transformed to a data matrix used to fit the classifier.

1. Text from username and short description
   (a) Tf-idf weighted unigram and bigram counts
   (b) Counts: Punctuation, Emoji, latin characters, city names, province names
   (c) Count of government words: A list of words most often associated with government agencies.
   (d) Count of non-government words: A list of words that would not likely be associated with government accounts (such as "celebrity," "athlete," "NBA," "TV program," "newspaper," etc).
   (e) Character length
2. Verified account or company account (binary)

[8]Because classification models of text often have high-dimensional feature-space, they are well-suited to the support vector machine (SVM) classifier. High dimensional spaces are more likely to be approximately linearly separable, a necessary condition for a linear SVM to work (*Joachims*, 1998).

3. Number of followers, following, posts (count, normalized)

5.3.2 Labeling Government Accounts

To begin the labeling process, I labeled a random sample of 3000 Weibo accounts that were following or followed by likely government astroturfers in a subset of comments meeting the requirements of rules 1, 2 and 4 in section 5.2.4.[9] After the initial sample was labeled, the remainder of Weibo accounts were labeled in batches of 100 using active learning until 10,000 accounts were labeled. Each batch comprised the 50 accounts with the shortest euclidean distance from each side of the class-separating hyperplane. Distance to the class-separating hyperplane measures uncertainty, as the SVM algorithm attempts to find the hyperplane that maximizes the margin between observations from each class. This means that the points closest to the hyperplane are the most likely points to be misclassified by the SVM. This iterative labeling process reduces the amount of labeled data needed to achieve a high performing classifier (*Brinker*, 2003; *Liu*, 2004; *Schohn and Cohn*, 2000; *Tong and Koller*, 2002). This process also can be useful for classification problems with class imbalances as labeling enough positive observations (government accounts) is difficult given their relative scarcity in the population. It also facilitates an automatic way of discovering concepts that seem obvious to a human but are actually quite difficult for a computer to differentiate.[10] See the algorithm for this labeling procedure in the appendix.

[9]Identifying social network connections on Weibo took a great deal of computational resources and time as there is no way of using the official Weibo API to obtain network connections. Instead, I had to use a headless browser to manually crawl each account, a process that took several weeks.

[10]For example, the classifier had trouble with hospital Weibo accounts, as most of these accounts include a city, and are organized by divisions and bureaus like government agencies. After one round of active learning which included several hospital accounts, the classifier was able to correctly classify these accounts.

Table 5.1: *Results on Held-out Development Set*

Class	**Precision**	**Recall**	**F1**	**Support**
Non-Govt	0.98	0.92	0.95	903
Govt	0.83	0.96	0.89	381
Avg./Total	0.94	0.93	0.93	1284

5.3.3 Performance of Government Weibo Classifier

I fit a linear support vector machine (SVM) on labeled data using the features described in the previous section. I employed randomized hyperparameter search and 5-fold cross validation to choose an optimal set of parameters from predefined continuous and discrete distributions of parameter values (*Bergstra and Bengio*, 2012). I used the hyperparameter set achieving the highest cross-validated F-measure score on the held-out development set. Table 5.1 shows the performance of the final model using precision, recall, and F1 metrics.[11]

Table 5.2: *Counts and Proportions of Government Accounts*

Type	**Count**	**%**
Domestic Security	2057	43.79
Propaganda Organs	1211	25.78
Courts, Local Govts., Procuratorates	483	10.28
Communist Youth Leagues	257	5.47
Economic Development	151	3.21
Other	538	11.45

The following is a rough measure of the types of government connections (followers and following) of government astroturfers.

[11] Precision measures how accurate the guesses of the relevant class are. Recall measures how many of the relevant class were recalled. F1 is the harmonic mean of precision and recall. F1 is useful for hyperparameter tuning and is often chosen as the metric used to tune hyperparameters in grid search, randomized search, or Bayesian optimization of hyperparameters (*Snoek et al.*, 2012).

$$\text{Precision} = \frac{\text{true positive}}{\text{true positive} + \text{false positive}} \quad \text{Recall} = \frac{\text{true positive}}{\text{true positive} + \text{false negative}} \quad \text{F-Measure} = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

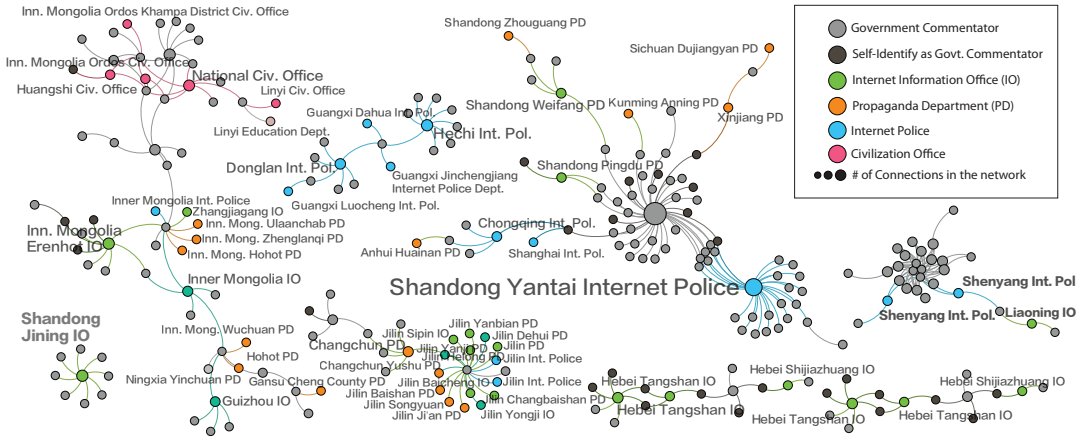
5.3.4 Network Structure and Inferring the Bureaucratic Affiliation of Astroturfers

The training set identified using rules in section 5.2.4 has network structures one would expect from documentary evidence describing the organization of these teams. It appears that government commentators cluster around Weibo accounts of similar region and bureaucratic type. This is consistent with government documents and manuals that suggest decentralization and specialization of commentating and opinion guidance work (*Miller*, 2016). As discussed in previous sections, government astroturfers are often connected to the social media accounts of affiliated bureaucracies. This means that astroturfers' social media data can be used to infer a astroturfer's bureaucratic affiliation. Though these guesses are crude, they confirm what is apparent in government documents and manuals, that the practice of government commentating is common across a wide range of bureaucracies in China.

Network graphs for propaganda organs can be seen in Figure 5.3, and network graphs for public security organs can be seen in Figure 5.4. Color nodes represent government accounts that are followed by at least one government astroturfer. Gray nodes represent government astroturfers. Several commentators in these figures even include their official title of "Internet Commentator" in their Weibo account bio. All government accounts are hand labeled according to bureaucracy type. A frequency table of government accounts by type can be seen in Table 5.2. A plurality of government astroturfers are part of China's burgeoning domestic security apparatus[12] Notably, propaganda departments do not make up the bulk of commentary, as is widely assumed to be true.

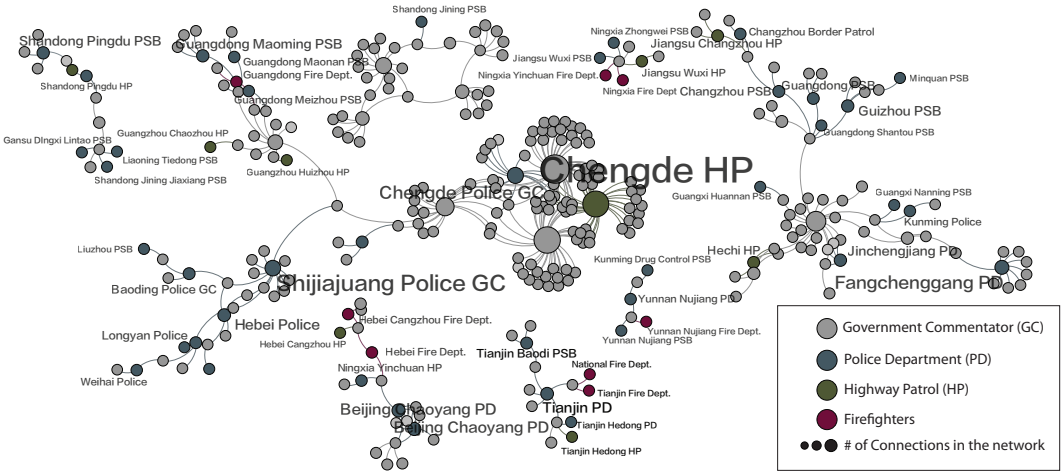
[12]In 2011, China's spending on domestic security outstripped it's military spending, and has continued to grow significantly year-on-year (*Buckley*, 2011).

Figure 5.3: Network Structure of Predicted Propaganda Accounts



Astroturfer accounts (gray) from the training set connected to propaganda bureaucracies' Weibo accounts (color).

Figure 5.4: Network Structure of Predicted Domestic Security Accounts



Astroturfer accounts (gray) from the training set connected to domestic security bureaucracies' Weibo accounts (color).

5.4 Government Astroturfer Classifier

With the training data discovered using the metadata search process outlined above, I trained a classifier to identify additional comments that could not be identified using the search procedure due to insufficient metadata. For the negative class, I randomly sampled an equal number of comments from the pool of comments one

would expect from an ordinary/civilian astroturfer, that is, posts where there is only one username associated with an IP address, and less than 5 comments per IP address. This made up the "non-government" class of comments. Once I fit classifiers on this training data, I could use them to retrieve probable astroturfer comments that were not identified using the search procedure due to missing metadata. I then validated classifiers on a sample of government astroturfer posts from the leaked Zhanggong Propaganda Department emails, showing that each classifier is predictive of ground-truth astroturfer comments.

5.4.1 Text Features

Below are the text features used in my classification model. To prevent overfitting and favor model sparsity, a subset of features are automatically discarded using the coefficients of a linear SVM with ℓ_1 norm penalty. Features with coefficient values lower than the mean of all coefficients are discarded (*Rakotomamonjy*, 2003).

Text classification using Chinese text requires an additional step due to the Chinese writing system's lack of spaces delineating the beginning and end of a word. Before using a word features, it is necessary to segment text. For this I use a hidden Markov model (HMM) to segment Chinese text.[13] After segmentation, stopwords[14] are removed.[15]

A full overview of features, including several custom features that proved discriminating are listed here:

1. Tf-idf[16] weighted unigram and bigram counts
2. Count of province names

[13]For this, I use `jieba`, a Python tool for Chinese text parsing: https://github.com/fxsjy/jieba

[14]Stopwords are words that carry little information (the, and, so, etc.)

[15]I use a stopword list that is a slightly modified version of the list used by Baidu, China's Google, for its natural language modeling.

[16]tf-idf stands for "term frequency inverse document frequency." The term frequency tf is the count of how many times a word appears in a document divided by the number of words in the document, the inverse document frequency idf measures how often a term occurs across all documents and is measured by $log(\frac{\text{total document count}}{\text{count of documents containing word}})$, and the tf-idf weight is their product $(tf \cdot idf)$ (*Lan et al.*, 2005)

3. Count of country names
4. Count of city names
5. Count of punctuation
6. Emoji count
7. Text length (in characters)

Table 5.3: *Results on Held-out Development Set*

Algorithm	**Acc.**	**Avg. Prec.**	**Avg. Rec.**	**Avg. F1**	**ZG Acc.**	**P(GC)**	**P(GC) s.e.**
SVM	0.6764	0.68	0.68	0.68	0.9050	0.1582	0.000072
LR	0.6779	0.68	0.68	0.68	0.9144	0.1710	0.000095
SVM, LR	0.6821	**0.69**	0.68	0.68	**0.9413**	0.1453	0.000353

5.4.2 Performance of Classifiers

Using the feature matrix outlined above, I fit several different classification algorithms to the training set consisting of observations that meet all the behavioral rules outlined in section 5.2.4.

As I did with the government Weibo classifier, I used 5-fold cross-validation and randomized hyperparameter search to select each model's parameters, optimizing for F-measure from predictions on held-out data (as explained above). After I tried a handful of classification algorithms and ensembles, I chose the 3 best performing models according to the average F1 score for both classes: a support vector machine (SVM), a logistic regression classifier (LR), and a majority vote ensemble classifier comprised of a SVM classifier and a logistic regression classifier. The performance of each model is well above the .5 baseline for binary classification problems and achieves an accuracy of .9 or above on the leaked Zhanggong comments.

5.4.3 Estimated Percent of Government Astroturfers in News Comment Sections

After fitting several classifiers to training data, I used them to predict the proportion of astroturfing commentary in the entire dataset. I use parametric bootstrap,

drawing 1000 random samples of $n = 25000$ to estimate the proportion of government comments. The raw proportion $\frac{\sum_{i}^{n} \hat{c}_i}{n}$ is biased because misclassification error is not uniform across classes. To adjust for this bias, we must estimate the proportion of government comments using the following equation, where $\hat{c} = 1$ if the model predicts the government comment class, and $\hat{c} = 0$ if the model predicts the non-government comment class:

$$P(c_i = 1) = \frac{P(\hat{c}_i = 1) - (1 - \text{recall}_0)}{\text{recall}_1 - (1 - \text{recall}_0)} \quad ; \quad \text{recall}_c = \frac{\text{true positive}_c}{\text{true positive}_c + \text{false negative}_c}$$

This estimator is unbiased so long as the performance metrics estimated with the training set "also hold in the unlabeled population set" (*Levy and KASS*, 1970; *Hopkins and King*, 2010). Though I do not assume a distribution for the estimates of $P(c_i = 1)$, the standard errors of the estimates in Table 5.3 assume normality, which appears reasonable as all distributions of resampled estimates appear unimodal and normal.

Based on the proportion estimates in Table 5.3, government comments appear to make up between 14.5%-17.3% of all commentary in the comment sections of the sample of news sources represented in my dataset.

Conclusion

I outline a method for retrieval of government astroturfers using non-text metadata. Based on careful and broad reading of government documents on the practice, I create rules to automatically label a training set, each of which is validated using ground truth data from the Zhanggong Propaganda Department emails. Using adjusted proportions from several different classifiers, run on two datasets, I show that between 14.5%-17.1% of all commentary represented in my dataset (approximately 6.75 million comments) come from government astroturfers. All classifiers predict

leaked propaganda data with at least 90% accuracy.

APPENDIX

APPENDIX A

A.1 Chapter 4

A.1.1 Intercoder Reliability

These are the results comparing the predictive power of coder 1 to coder 2's decisions using the area under the receiver operator curve (ROC AUC). ROC AUC is a better measure than raw accuracy because it is not misleading in situations where there are class imbalances. Nearly all categories are above the acceptable measure of .7 and most are above .85, which is highly reliable.

Table A.1: Intercoder Reliability Measures

Category	**AUC**	**Category**	**AUC**	**Category**	**AUC**
Col. Action	0.89	Government	0.82	Reoccurring Political Event	0.85
Social Groups	0.87	Central Government	0.78	Sensitive Anniversary	0.89
Petitions	0.8	State-Owned Enterprise	0.63	Regular Political Event	0.67
Protest	0.86	Government Policy	0.72	Sexuality	0.82
Social Activism	0.85	Political Humor	0.96	Pornography	0.51
Strikes	1	Government Leadership	0.92	Sina	0.94
Commercial	0.81	Local Government	0.73	Sina Censorship	0.98
Sina's Competitors	0.51	Party Ideology	0.83	Sina Company Business	0.93
Corruption	0.93	Leadership Personal Information	0.7	Hong Kong, Taiwan, Macau	0.98
Central Govt. Corruption	0.91	Xi Jinping	0.98	Hong Kong, Macau	0.99
Local Govt. Corruption	0.84	Government Criticism	0.7	Taiwan	0.99
Crime	0.81	Censorship Policy	0.58	Ethnicity	0.98
Financial Crime	0.89	Power Struggle	0.59	Tibetan	0.94
Political Crime	0.86	Nationalism	0.79	Uighur	0.96
Violent Crime	0.84	Party History	0.67	Other Ethnic Minority	0.94
Illegal Goods/Services	0.89	Military	0.78	Entertainment	0.73
Disaster	0.91	Territorial Disputes	0.94	Terrorism	0.99
Man-made Disaster	0.9	Rumors	0.91		
Natural Disaster	0.76	Non-political Rumors	0.98		
Foreign Media	0.81	Political Rumors	0.88		

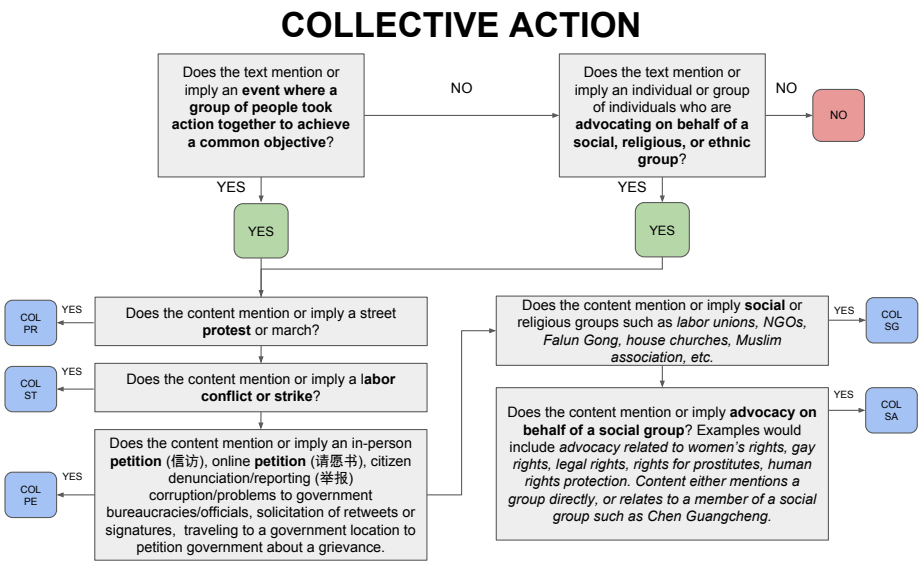
A.1.2 Coding Diagrams

The coding diagrams below were used by research assistants during coding. Blue boxes represent secondary topic categories. Abbreviations for topic categories can be found in Table A.2. A text version of the coding scheme is included below each flow chart.

Table A.2: Abbreviation Mapping

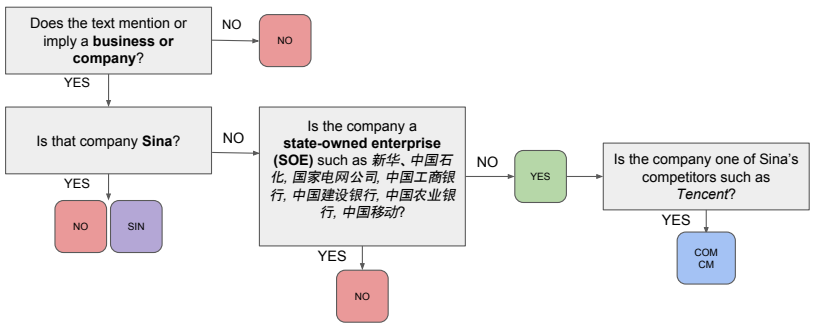
Abbreviation	**Category Name**	**Abbreviation**	**Category Name**
COL_PE	Petition	GOV_CP	Censorship Policy
COL_PR	Protest	GOV_CR	Criticism
COL_SA	Social Activism	GOV_GP	Government Policy
COL_SG	Social Groups	GOV_HU	Humor, Satire
COL_ST	Strike/Labor Disputes	GOV_LE	Government Leadership
COM_CM	Competitors	GOV_LO	Local/Subnational Government
COR_CE	Central/National Government	GOV_PA	Party Ideology
COR_LO	Local/Subnational Government	GOV_PE	Personal Information
CRI_CC	Cyber Crime	GOV_PS	Power Struggle
CRI_FI	Financial Crime	GOV_XI	Xi Jinping
CRI_GM	Gambling	HKT_HK	Hong Kong, Macau, and Taiwan
CRI_IG	Illicit Goods and Services	HKT_TW	Taiwan
CRI_PO	Police	NAT_HI	History
CRI_VI	Violent Crime	NAT_MI	Military
DIS_MA	Man-made Disaster	NAT_TE	Territorial disputes
DIS_NA	Natural Disaster	RUM_NO	Non-political Rumors
ENT_	Entertainment	RUM_POL	Political Rumor
ETH_OT	Other Ethnic Group	SEN_AN	Anniversary
ETH_TI	Tibetan	SEN_GO	Government Business
ETH_UI	Uighur	SEX_LG	LGBT
FOR_FOR	Foreign Media	SEX_POR	Pornography
GOV_CE	Central/National Government	SEX_SE	Sexually Suggestive Content
GOV_CO	State-Owned Enterprise	TER_	Terrorism

Figure A.1: Collective Action Coding Diagram



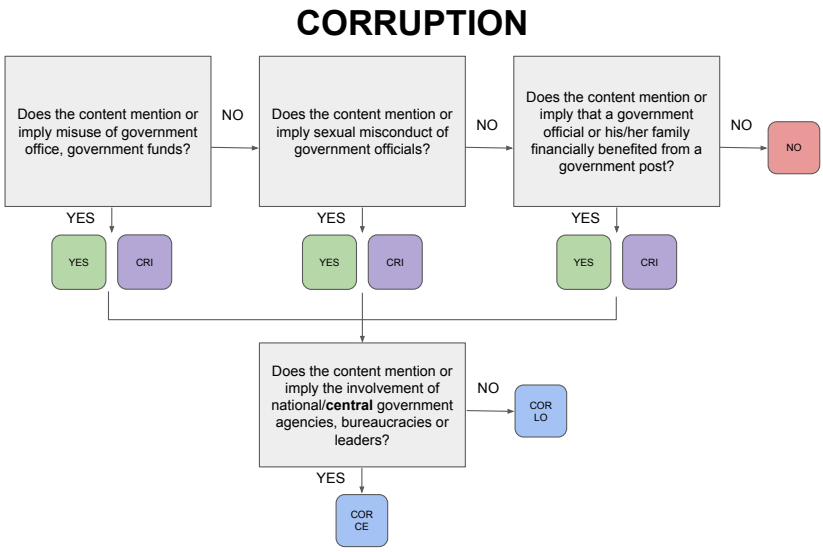
The content either 1) mentions or implies an event where a group of people took action together to achieve a common objective, or 2) mentions or implies an individual or group of individuals who are advocating on behalf of a social, religious, or ethnic group.

Figure A.2: Commercial Coding Diagram



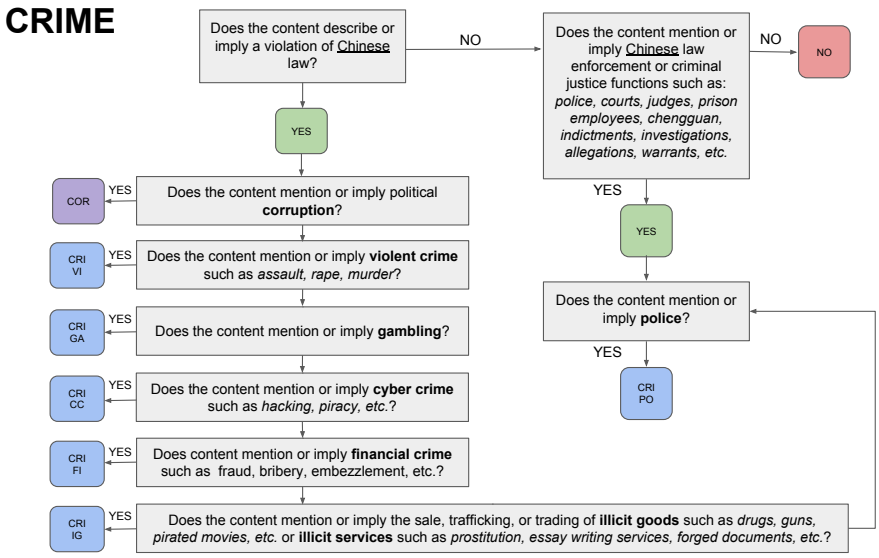
The text mentions or implies a business or company that is not Sina and is not a state-owned enterprise (SOE).

Figure A.3: Corruption Coding Diagram



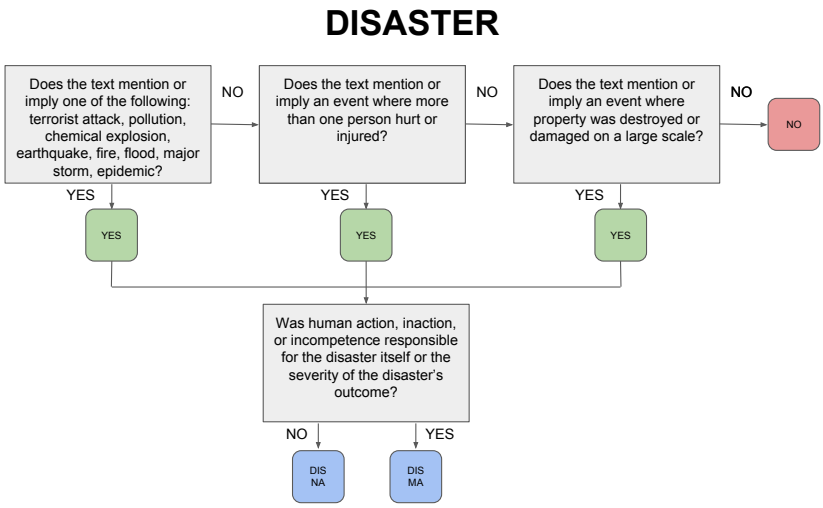
The content mentions or implies any of the following: 1) misuse of local government office or local government funds, 2) sexual misconduct of local government officials 3) a local government official and/or his/her family financially benefiting from a government post.

Figure A.4: Crime Coding Diagram



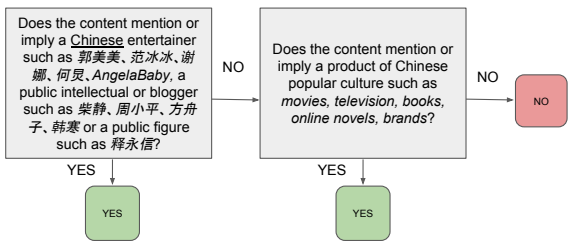
The content mentions or implies either 1) a violation of Chinese law, 2) Chinese law enforcement or criminal justice functions such as: police, courts, judges, prison employees, chengguan, indictments, investigations, allegations, warrants, etc.

Figure A.5: Disaster Coding Diagram



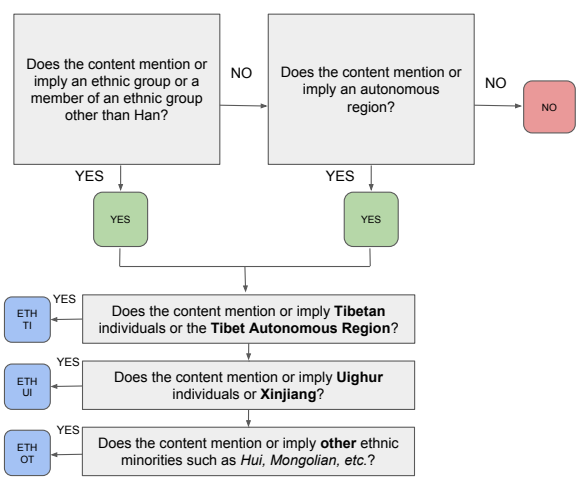
The content mentions or implies one of the following: 1) an event where more than one person hurt or injured, 2) an event where property was destroyed or damaged on a large scale, 3) any of the following: terrorist attack, pollution, chemical explosion, earthquake, fire, flood, major storm, epidemic.

Figure A.6: Entertainment Coding Diagram



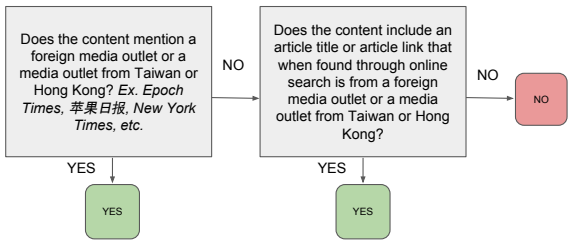
The content mentions or implies one of the following: 1) a Chinese entertainer, public intellectual, blogger, or a public figure, 2) a product of Chinese popular culture such as movies, television, books, online novels, brands.

Figure A.7: Ethnicity Coding Diagram



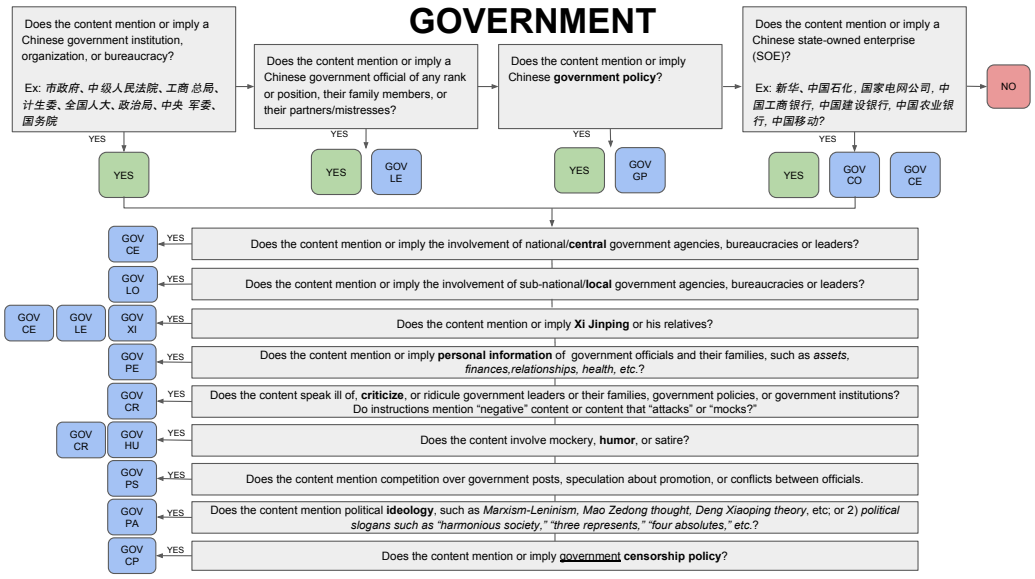
The content mentions or implies one of the following: 1) an ethnic group or a member of an ethnic group other than Han, 2) an autonomous region in China.

Figure A.8: Foreign Media Coding Diagram



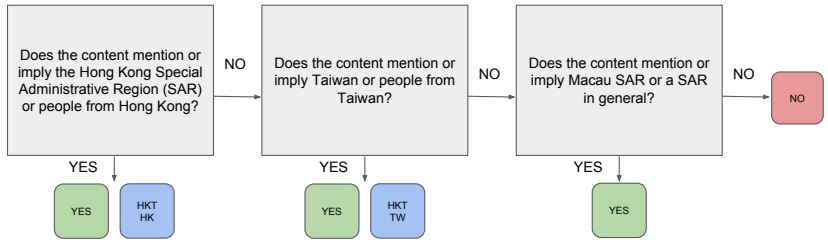
The content mentions a foreign media outlet or a media outlet from Taiwan or Hong Kong. This includes content with an article title or article link that when searched is from a foreign media outlet or a media outlet from Taiwan or Hong Kong.

Figure A.9: Government Coding Diagram



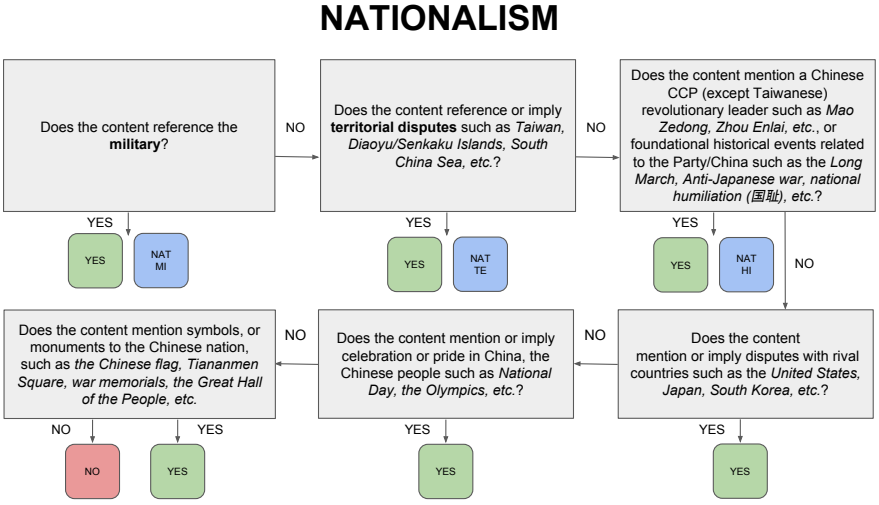
The content mentions or implies a Chinese government institution, organization, or bureaucracy, a Chinese government official of any rank or position, their family members or their partners/mistresses, a Chinese government policy, or a Chinese state-owned enterprise (SOE).

Figure A.10: Hong Kong/Taiwan Coding Diagram



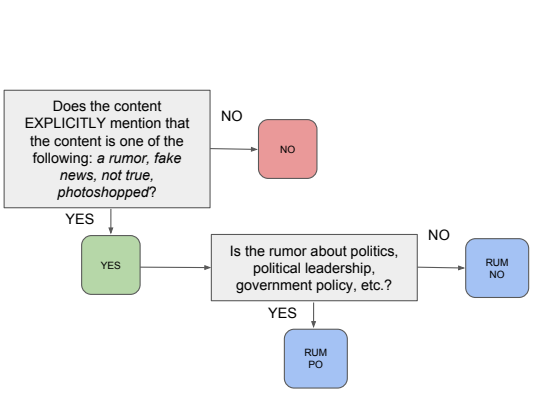
The content mentions or implies any of the following: 1) the Hong Kong Special Administrative Region (SAR) or people from Hong Kong, 2) Taiwan or people from Taiwan, 3) Macau SAR, an SAR in general, or people from an SAR.

Figure A.11: Nationalism Coding Diagram



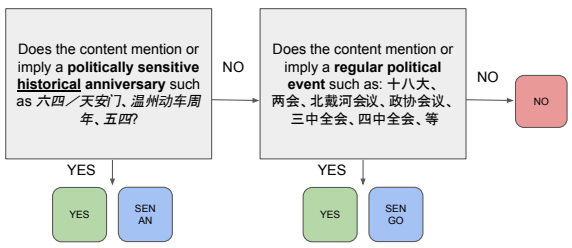
The content references or implies any of the following: 1) territorial disputes such as Taiwan, Diaoyu/Senkaku Islands, South China Sea, etc., 2) a Chinese CCP (except Taiwanese) revolutionary leader such as Mao Zedong, Zhou Enlai, etc., or foundational historical events related to the Party/China such as the Long March, Anti-Japanese war, national humiliation (国耻), etc., 3) a Chinese CCP (except Taiwanese) revolutionary leader such as Mao Zedong, Zhou Enlai, etc., or foundational historical events related to the Party/China such as the Long March, Anti-Japanese war, national humiliation (国耻), etc., 4) celebration or pride in China, the Chinese people such as National Day, the Olympics, etc., 5) disputes with rival countries such as the United States, Japan, South Korea, etc.

Figure A.12: Rumors Coding Diagram



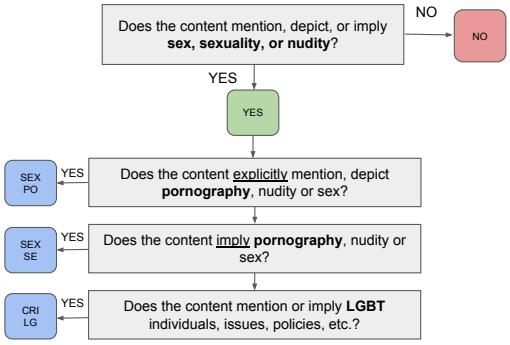
The content EXPLICITLY mentions that the content is one of the following: a rumor, fake news, not true, photoshopped.

Figure A.13: Sensitive Anniversary Coding Diagram



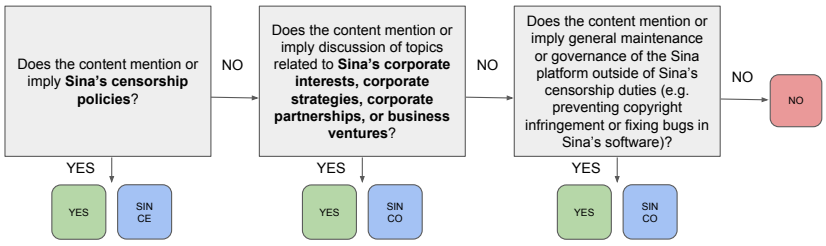
The content mentions or implies any of the following: 1) a politically sensitive historical anniversary such as June 4 or the Wenzhou train crash, 2) a regular political event such as a party congress, the two meetings, etc.

Figure A.14: Sexuality Coding Diagram



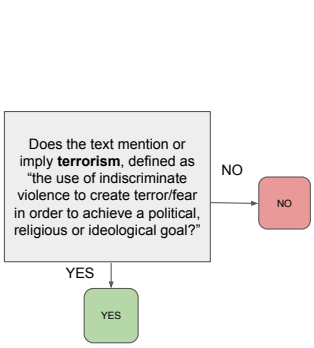
The content mentions, depicts, or implies sex, sexuality, or nudity.

Figure A.15: Sina Coding Diagram



The content mentions or implies one of the following: 1) Sina's censorship policies, 2) discussion of topics related to Sina's corporate interests, corporate strategies, corporate partnerships, or business ventures, 3) general maintenance or governance of the Sina platform outside of Sina's censorship duties (e.g. preventing copyright infringement or fixing bugs in Sina's software)

Figure A.16: Terrorism Coding Diagram

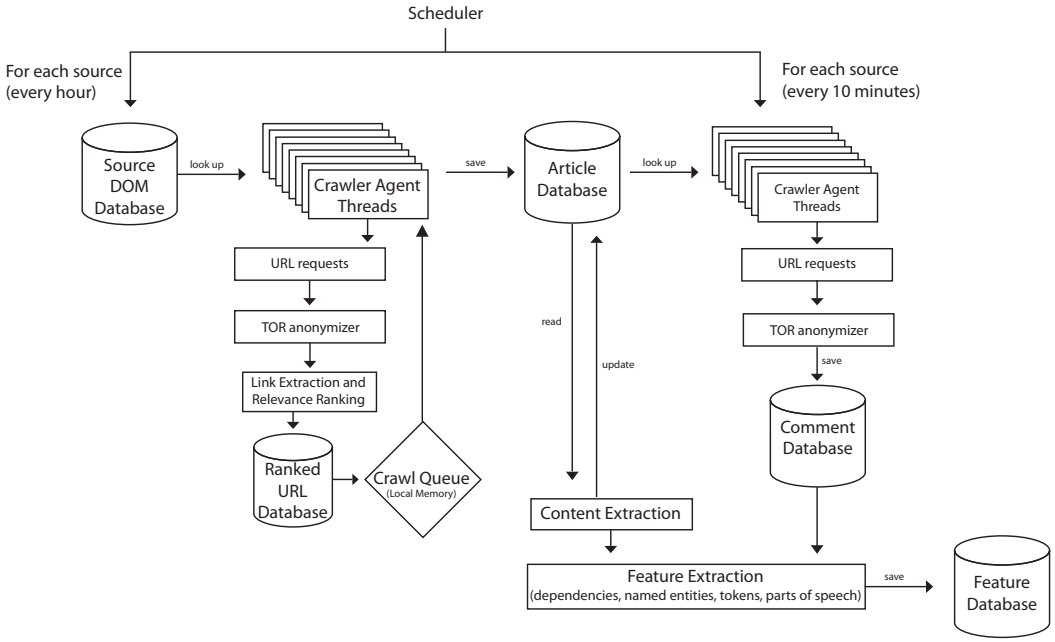


The text mentions or implies terrorism, defined as "the use of indiscriminate violence to create terror/fear in order to achieve a political, religious or ideological goal?"

A.2 Chapter 5

A.2.1 Data Collection

Figure A.17: Data Collection and Processing Architecture



I have built software to collect comment and article data at 10 minute intervals persistently on a dedicated 64-core cluster, gathering approximately 2.5 million new comments and 75,000 articles every week. Data are collected via requests to news websites through the TOR network[1] which obfuscates the server's IP address to prevent request throttling. Data have been collected since late 2015, spanning just over a year's time. The scraping software targets 19 popular news outlets that vary in their state-affiliation, level of commercialization, and region. Unless articles and posts are censored, or are deliberately hidden, I collect the population of articles matching a site-specific regular expression for news articles from each of these news websites.[2]

[1] https://www.torproject.org/

[2] Article text and metadata along with each article's comments and related user metadata are collected using API's I uncovered hidden in several sites' software architectures, or slower, less efficient headless browser scrapers that gather data from the DOM of a website.

In total, I have collected approximately 6 million articles and 70 million comments. Text features and data are processed asynchronously. Please see Figure A.17 for a detailed diagram of data collection and processing architecture.

A.2.2 Typology of Online Commentary

Along the "organization" dimension, I specify two categories: "institutional" and "individual." These categories describe the behavior I observe at each IP address. An IP address represents a unique local internet network. This can be a wireless network for a home, office, or coffee shop. At a normal IP address, one would expect an individual or groups of individuals to behave in certain ways. For example, users at a coffee shop are likely to visit a wide range of websites and it would be very unlikely that two individuals at a single coffee shop would comment on the same news article. One would expect there to be a great degree of variance in the topics of social media posts and articles upon which individuals from this IP comment. Network activity such as this indicates a lack of group coordination and an "organization" level that matches uncoordinated, or "individual" browsing behavior. Conversely, one might observe an IP address where several users visit the same news article and post similar content in favor of a government policy, a product, or a church event. This type of activity seems more characteristic of a PR firm's office, or a government bureaucracy. If one observed messages from a single IP address that were coordinated in this way, one might assume that the level of organization of comments from this IP address is "institutional." Along the "alignment" dimension I specify three categories: "aligned", "neutral", and "opposed." These levels are fairly straightforward. Commentary that is "aligned" with state interests is political in nature, and is consistent with the Party's ideology. Commentary that is "opposed" to state interests is also political in nature, but criticizes the Party's policies, or expresses opinions that are in opposed to the interests of the regime. Naturally, "neutral" messages are not political in nature, and thus do not lean one way or another, an example would be news about a basketball game. Categories along this "Alignment of Content to State Interests" dimension help distinguish if comments within a single IP address are consistent with messages we would expect to come from

state actors. This dimension helps us disambiguate regime, corporate and foreign sources of "institutional" commentary to aid in accurate detection of government astroturfing with which we can train a text classification model.

"Government astroturfer commentary", which is represented in the typology by the cell in the top left, represents pro-regime commentary that comes from regime actors. I define regime actors as any person employed by a government agency (i.e. bureaucrats, politicians, government contractors). The motivation of this commentary is either to "guide opinion" in the direction of state interests, signal regime strength, or organize the the masses (i.e. the "mass line"[3]).

"Corporate-sponsored commentary" is commentary that is made for advertisement or PR purposes, in support of a corporation or non-political organization. This encompasses posts by individuals belonging to the "Water Army," the corporate counterparts of the "Fifty Cent Party," who systematically post positive comments about a business or product.[4] The motivation of this commentary is to influence how individuals spend their money (i.e. to advertise).

"Foreign-regime sponsored commentary," is commentary that is systematically posted by foreign governments. Many in China believe that foreign governments are also involved in systematically posting pro-West and pro-democracy comments on Chinese forums. Even if foreign governments do not engage in this behavior, the mere idea of it is meaningful because it clarifies how the Chinese government rationalizes the need for government astroturfers as a battle against foreign incursion into

[3]The mass line (群众路线) is a theoretical leadership method that was developed in the Chinese revolution and is now a part of Chinese Marxist-Leninist communication theory. The mass line stresses that propaganda is a state tool for organizing the masses. According to these theories, the responsibility of state leaders is to systematize diffuse thoughts, beliefs, and opinions of citizens.

[4]The term corporate is broadly defined, and can also include religious proselytization, consistent with religion and politics literature describing the "firm-like" structure of religious institutions (*Gill*, 2008)

the sovereign space of the Chinese internet. The motivation of this commentary is to guide opinion in the direction of a foreign state's interests, or to persuade individuals to adopt anti-government attitudes within their own country. Though little evidence of these actors exists in China, other states, such as Egypt may experience more foreign-regime sponsored commentary from Israel or psuedo-state actors like ISIS.

Organic commentary (OC) represents commentary that one would expect to see from ordinary individuals. This commentary is unsystematic, and is not aimed at changing the nature of discussions. This type of commentary is represented in blue in Figure 5.1. The motivation of this commentary is simply to participate in discussions of content. Motivations may differ from individual to individual, but the average person does not usually have any systematized agenda behind their comments, and even if they do have an agenda (such as evangelizing or trolling), they act alone.

BIBLIOGRAPHY

BIBLIOGRAPHY

Ai, W. (2012), China's paid trolls: Meet the 50-cent party, *The New Statesman*.

Ang, Y. Y. (2014), Authoritarian restraints on online activism revisited: Why "i-paid-a-bribe" worked in india but failed in china, *Comparative Politics*, *47*(1), 21–40.

Ang, Y. Y. (2016), *How China escaped the poverty trap*, Cornell University Press.

Anonymous (2014), 2014 guizhou communist youth league online commentator team construction work instructions.

Bergstra, J., and Y. Bengio (2012), Random search for hyper-parameter optimization, *Journal of Machine Learning Research*, *13*(Feb), 281–305.

Brady, A.-M. (2009), *Marketing dictatorship: Propaganda and thought work in contemporary China*, Rowman & Littlefield.

Brinker, K. (2003), Incorporating diversity in active learning with support vector machines, in *ICML*, vol. 3, pp. 59–66.

Buckley, C. (2011), China internal security spending jumps past army budget.

Cai, Y. (2008), Local governments and the suppression of popular resistance in china, *The China Quarterly*, *193*, 24–42.

Cairns, C. (2016a), Fragmented authoritarianism? reforms to china's internet censorship system under xi jinping, Ph.D. thesis, Cornell University.

Cairns, C. (2016b), The 'social media shock', unpublished dissertation chapter.

Cairns, C. (2017), China's weibo experiment: Social media (non-) censorship and autocratic responsiveness, Ph.D. thesis, Cornell University, unpublished dissertation.

Calvert, R. L., M. D. McCubbins, and B. R. Weingast (1989), A theory of political control and agency discretion, *American journal of political science*, pp. 588–611.

Capoccia, G., and D. Ziblatt (2010), The historical turn in democratization studies: A new research agenda for europe and beyond.

CCP (2013), Guanyu jin yi bu jiaqiang fanfuchanglian sucai bao song gongzuo de yijian.

Chang, J., and J. Halliday (2005), Mao: The untold story, *London, Jonathan Cape*.

Chen, J., and Y. Xu (2016), Why do authoritarian regimes allow citizens to voice opinions publicly?, *The Journal of Politics, Forthcoming*.

Chen, X. (2017), Origins of informal coercion in china, *Politics & Society*, *45*(1), 67–89, doi:10.1177/0032329216681489.

Cho, C. H., M. L. Martens, H. Kim, and M. Rodrigue (2011), Astroturfing global warming: It isn't always greener on the other side of the fence, *Journal of Business Ethics*, *104*(4), 571–587.

Crandall, J. R., M. Crete-Nishihata, J. Knockel, S. McKune, A. Senft, D. Tseng, and G. Wiseman (2013), Chat program censorship and surveillance in china: Tracking tom-skype and sina uc, *First Monday*, *18*(7).

Crete-Nishihata, M., A. Hilts, J. Knockel, J. Q. Ng, L. Ruan, and G. Wiseman (2016), Harmonized histories? a year of fragmented censorship across chinese live streaming applications.

Crete-Nishihata, M., J. Knockel, B. Miller, J. Q. Ng, L. Ruan, and R. Xiong (2017), Remembering liu xiaobo: Analyzing censorship of the death of liu xiaobo on wechat and weibo.

Cribben, I., A. Esarey, R. Han, and X. Qiang (2018), Central and local preferences for information control in china, unpublished paper draft.

Crouch, C. (2004), *Post-democracy*, Polity Cambridge.

Deng, Y., and K. J. O'Brien (2013), Relational repression in china: using social ties to demobilize protesters, *The China Quarterly*, *215*, 533–552.

Diamond, L. (2010), Liberation technology, *Journal of Democracy*, *21*(3), 69–83.

Dickson, B. (2016), *The Dictator's Dilemma: The Chinese Communist Party's Strategy for Survival*, Oxford University Press.

Dimitrov, M. K. (2017), The political logic of media control in china, *Problems of Post-Communism*, *64*(3-4), 121–127.

Dimitrov, M. K., et al. (2013), Understanding communist collapse and resilience, *Why communism did not collapse: Understanding authoritarian regime resilience in Asia and Europe*, pp. 3–39.

Egorov, G., S. Guriev, and K. Sonin (2009a), Media freedom in dictatorships, *Unpublished paper, Yale University*.

Egorov, G., S. Guriev, and K. Sonin (2009b), Why resource-poor dictators allow freer media: A theory and evidence from panel data, *American political science Review*, *103*(04), 645–668.

Esarey, A., and Q. Xiao (2011), Digital communication and political change in china, *International Journal of Communication*, *5*, 22.

Fu, K.-w., and M. Chau (2013), Reality check for the chinese microblog space: a random sampling approach, *PloS one*, *8*(3), e58,356.

Gallagher, M., and B. Miller (2018), Legitimation and control: Social media governance in china, unpublished paper draft.

Gao, H., and Z. Zhang (2011), Wangluo yuqing yu shehui wending.

Gehlbach, S., and K. Sonin (2014), Government control of the media, *Journal of Public Economics*, *118*, 163–171.

Gill, A. (2008), *Rendering unto Caesar: the Catholic Church and the state in Latin America*, University of Chicago Press.

Global Times Editorial Team (2016), Sheping: Hafo tuandui dui suowei "wumaodang" yizhibanjue.

Gueorguiev, D. D., and E. J. Malesky (2018), Revisiting selective censorship in china, unpublished paper draft.

Guriev, S. M., and D. Treisman (2015), How modern dictators survive: Cooptation, censorship, propaganda, and repression.

Han, R. (2015a), Defending the authoritarian regime online: China's "voluntary fifty-cent army", *The China Quarterly*, pp. 1–20.

Han, R. (2015b), Manufacturing consent in cyberspace: China's "fifty-cent army", *Journal of Current Chinese Affairs*, *44*(2), 105–134.

Han, R. (2018), *Contesting cyberspace in China: Online expression and authoritarian resilience*, Columbia University Press.

Heilmann, S. (2008), From local experiments to national policy: The origins of china's distinctive policy process, *The China Journal*, (59), 1–30.

Hintz, A. (2016), Restricting digital sites of dissent: commercial social media and free expression, *Critical Discourse Studies*, *13*(3), 325–340.

Hopkins, D. J., and G. King (2010), A method of automated nonparametric content analysis for social science, *American Journal of Political Science*, *54*(1), 229–247.

Huang, C., and K. Zhai (2013), Xi jinping rallies party for propaganda war on internet.

Jin, H., Y. Qian, and B. R. Weingast (2005), Regional decentralization and fiscal incentives: Federalism, chinese style, *Journal of public economics*, *89*(9), 1719–1742.

Joachims, T. (1998), Text categorization with support vector machines: Learning with many relevant features, in *European conference on machine learning*, pp. 137–142, Springer.

Keane, M. (2001), Broadcasting policy, creative compliance and the myth of civil society in china, *Media, Culture & Society*, *23*(6), 783–798.

Keller, F. B., D. Schoch, S. Stier, and J. Yang (2017), How to manipulate social media: Analyzing political astroturfing using ground truth data from south korea., in *ICWSM*, pp. 564–567.

King, G., J. Pan, and M. E. Roberts (2013), How censorship in china allows government criticism but silences collective expression, *American Political Science Review*, *107*(02), 326–343.

King, G., J. Pan, and M. E. Roberts (2014), Reverse-engineering censorship in china: Randomized experimentation and participant observation, *Science*, *345*(6199), 1251,722.

King, G., J. Pan, and M. E. Roberts (2016), How the chinese government fabricates social media posts for strategic distraction, not engaged argument, unpublished paper draft.

Knockel, J., J. R. Crandall, and J. Saia (2011), Three researchers, five conjectures: An empirical analysis of tom-skype censorship and surveillance., in *FOCI*.

Knockel, J., M. Crete-Nishihata, J. Q. Ng, A. Senft, and J. R. Crandall (2015), Every rose has its thorn: Censorship and surveillance on social video platforms in china, in *5th USENIX Workshop on Free and Open Communications on the Internet (FOCI 15)*.

Knockel, J., L. Ruan, and M. Crete-Nishihata (2017), Measuring decentralization of chinese keyword censorship via mobile games, in *7th USENIX Workshop on Free and Open Communications on the Internet (FOCI 17)*, USENIX Association, Vancouver, BC.

Kuran, T. (1987), Preference falsification, policy continuity and collective conservatism, *The Economic Journal*, *97*(387), 642–665.

Lan, M., C.-L. Tan, H.-B. Low, and S.-Y. Sung (2005), A comprehensive comparative study on term weighting schemes for text categorization with support vector machines, in *Special interest tracks and posters of the 14th international conference on World Wide Web*, pp. 1032–1033, ACM.

Lei, Y.-W. (2017), *The contentious public sphere: Law, media, and authoritarian rule in China*, Princeton University Press.

Lessig, L. (1999), Code is law, *The Industry Standard*, *18*.

Levy, P. S., and E. H. KASS (1970), A three-population model for sequential screening for bacteriuria, *American Journal of Epidemiology*, *91*(2), 148–154.

Lieberthal, K. (1995), *Governing China: From revolution through reform*, WW Norton.

Liu, L., and B. R. Weingast (2017), Taobao, federalism, and the emergence of law, chinese style.

Liu, Y. (2004), Active learning with support vector machine applied to gene expression data for cancer classification, *Journal of chemical information and computer sciences*, *44*(6), 1936–1941.

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

Lynch, D. C. (1999), *After the propaganda state: Media, politics, and" thought work" in reformed China*, Stanford University Press.

MacKinnon, R. (2009), China's censorship 2.0: How companies censor bloggers, *First Monday*, *14*(2).

Mai, B. (2016), Data-driven surveillance as a business: Analysis of government expenditure on online public opinion monitoring, unpublished paper draft.

Malesky, E., and P. Schuler (2011), The single-party dictator's dilemma: Information in elections without opposition, *Legislative Studies Quarterly*, *36*(4), 491–530.

Mertha, A. (2009), "fragmented authoritarianism 2.0": political pluralization in the chinese policy process, *The China Quarterly*, *200*, 995–1012.

Miller, B. (2016), Surveillance-driven authoritarian learning from "public opinion emergencies" in china, unpublished paper draft.

Miller, B. (2018), The limits of commercialized censorship in china, unpublished paper draft.

Montinola, G., Y. Qian, and B. R. Weingast (1995), Federalism, chinese style: the political basis for economic success in china, *World politics*, *48*(01), 50–81.

Morozov, E. (2012), *The net delusion: The dark side of Internet freedom*, PublicAffairs.

Ng, J. Q. (2016), Politics, rumors, and ambiguity: Tracking censorship on wechat's public accounts platform, *citizenlab.org*.

O'Brien, K. J. (1994), Implementing political reform in china's villages, *The Australian Journal of Chinese Affairs*, (32), 33–59.

O'Brien, K. J., and Y. Deng (2015), Repression backfires: tactical radicalization and protest spectacle in rural china, *Journal of Contemporary China*, *24*(93), 457–470.

Oksenberg, M., and K. G. Lieberthal (1988), Policy making in china: Leaders, structures, and processes.

Ong, L. H. (2015), 'thugs-for-hire': State coercion and everyday repression in china.

Pan, J. (2016), How market dynamics of domestic and foreign social media firms shape strategies of internet censorship, *Problems of Post-Communism*.

Peoples Daily Weibo Data Center (2015), 2015 nian yi jidu renmin ribao zhengwu zhishu weibo yingxiangli baogao.

PRC State Council General Office (2016), Guowuyuan bangongting guanyu zai zhengwu gongkai gongzuo zhong jin yibu zuohao zhengwu yuqing huiying de tongzhi.

Rakotomamonjy, A. (2003), Variable selection using svm-based criteria, *Journal of machine learning research*, *3*(Mar), 1357–1370.

Roberts, M. (2015), Experiencing censorship emboldens internet users and decreases government support in china, *Unpublished Working Paper, URL http://www.margaretroberts.net/wp-content/uploads/2015/07/fear.pdf*.

Roberts, M. (2017), *The Censorship Tax Information Distortion Within China's Great Firewall*, Princeton University Press.

Roberts, M. E. (2018), *Censored: Distraction and Diversion Inside Chinas Great Firewall*, Princeton University Press.

Ruan, L., J. Knockel, J. Q. Ng, and M. Crete-Nishihata (2016), One app, two systems: How wechat uses one censorship policy in china and another internationally, *citizenlab.org*.

Rundlett, A., and M. W. Svolik (2016), Deliver the vote! micromotives and macrobehavior in electoral fraud, *American Political Science Review*, *110*(1), 180–197.

Schohn, G., and D. Cohn (2000), Less is more: Active learning with support vector machines, in *ICML*, pp. 839–846, Citeseer.

Schurmann, F. (1966), *Ideology and organization in communist China*, Univ of California Press.

Shambaugh, D. (2007), China's propaganda system: Institutions, processes and efficacy, *The China Journal*, (57), 25–58.

Smith, T. F., and M. S. Waterman (1981), Comparison of biosequences, *Advances in applied mathematics*, *2*(4), 482–489.

Snoek, J., H. Larochelle, and R. P. Adams (2012), Practical bayesian optimization of machine learning algorithms, in *Advances in neural information processing systems*, pp. 2951–2959.

Stern, R. E., and J. Hassid (2012), Amplifying silence: uncertainty and control parables in contemporary china, *Comparative Political Studies*, *45*(10), 1230–1254.

Stern, R. E., and K. J. O'Brien (2012), Politics at the boundary: Mixed signals and the chinese state, *Modern China*, *38*(2), 174–198.

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

Stockmann, D., and T. Luo (2017), Which social media facilitate online public opinion in china?, *Problems of Post-Communism*, pp. 1–14.

Stukal, D., S. Sanovich, and J. Tucker (2017), Detecting political bots on russian twitter, unpublished paper draft.

Tager, J., K. Glenn Bass, and S. Lopez (2017), Forbidden feeds: Government controls on social media in china.

Tanash, R. S., Z. Chen, T. Thakur, D. S. Wallach, and D. Subramanian (2015), Known unknowns: An analysis of twitter censorship in turkey, in *Proceedings of the 14th ACM Workshop on Privacy in the Electronic Society*, pp. 11–20, ACM.

Tong, S., and D. Koller (2002), Support vector machine active learning with applications to text classification, *The Journal of Machine Learning Research*, *2*, 45–66.

Truex, R. (2014), The returns to office in a "rubber stamp" parliament, *American Political Science Review*, *108*(2), 235–251.

Villeneuve, N. (2008), Search monitor: Toward a measure of transparency.

Volland, N. (2003), The control of the media in the people's republic of china, Ph.D. thesis.

Wagner, B. (2008), Modifying the data stream: Deep packet inspection and internet censorship.

Wagner, B. (2009), Deep packet inspection and internet censorship: International convergence on an'integrated technology of control', *Available at SSRN 2621410*.

Wallace, J. (2014), *Cities and stability: Urbanization, redistribution, and regime survival in China*, Oxford University Press.

Wang, Y. (2016a), The business of censorship: Documents show how weibo filters sensitive news in china.

Wang, Y. (2016b), Read and delete: How weibo's censors tackle dissent and free speech.

Wintrobe, R., et al. (1998), *The political economy of dictatorship*, vol. 6, Cambridge Univ Press.

Xinhua (2016), Xi jinping: Rang hulianwang geng hao zaofu guojia he renmin.

Yang, G. (2013), *The power of the Internet in China: Citizen activism online*, Columbia University Press.

Yang, X., Q. Yang, and C. Wilson (2015), Penny for your thoughts: Searching for the 50 cent party on sina weibo, in *Ninth International AAAI Conference on Web and Social Media*.

Zhang, S. (2011), Rang wumao wumeifen zaodian cheng lishi.

Zheng, Y. (2007), *De facto federalism in China: Reforms and dynamics of central-local relations*, vol. 7, World Scientific.

Zou, C. L., Hongqiang, and G. Su (2015), *Lingdao ganbu: Wangluo yuqing gongzuo zhinan Work Guide for Public Opinion*, People's Daily Press.