Journal of Contemporary China

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/cjcc20

Censorship in the Semi-private Domain: A Theory of Cross-domain Variation and Evidence from WeChat

Elliot Ji & Zack Bowersox

To cite this article: Elliot Ji & Zack Bowersox (2022) Censorship in the Semi-private Domain: A Theory of Cross-domain Variation and Evidence from WeChat, Journal of Contemporary China, 31:136, 592-608, DOI: 10.1080/10670564.2021.1985839

To link to this article: https://doi.org/10.1080/10670564.2021.1985839

JOURNAL OF CONTEMPORARY CHINA
2022, VOL. 31, NO. 136, 592–608
https://doi.org/10.1080/10670564.2021.1985839

Routledge
Taylor & Francis Group

Censorship in the Semi-private Domain: A Theory of Cross-domain Variation and Evidence from WeChat

Elliot Ji[a] and Zack Bowersox[b]

[a]Princeton University, USA; [b]Washington University in St. Louis, USA

ABSTRACT

Previous work addressing China's censorship regime has primarily focused on public information from social media sites and not information shared more intimately. This article focuses on semi-private information and its impact on collective action, using an original experiment to test established censorship theories in this overlooked domain. The results suggest that censors treat information critiquing the government and calling for collective action with equal hostility, unlike in the public domain in which the former category is more likely to be disregarded. Further, this article finds evidence of human involvement in semi-private domain censorship. This study aims to complement existing literature on authoritarian control of information with a view to the regime's effort to prevent collective action and political opportunities that can be exploited by dissent.

Introduction

Authoritarian regimes have been known to preference social stability, dissuade collective action, and actively prevent 'trigger events' from inciting mass protest.[1] The logic of such activities is largely attributed to the government's sense of security.[2] That is, if an executive or their administration feels that their job security, their tenure in office, is at risk, they are more likely to repress. Signs of insecurity can be increased protests, observable mobilization of the opposition party and an overall increase in civil discord. Acts of physical repression, such as political arrests, curfews, and the disappearing of opposition members, have been found by some scholars to have a preventative effect on collective action, thus reducing the overt signs of executive insecurity.[3]

Online censorship constitutes yet another type of repression, an invisible, mostly non-violent tactic that represses a citizen's right to information and expression. In the long run, the effort of censorship is geared toward impeding the people's ability to engage in collective action. One might suspect that information conveying grievances, discontent, and calls for assembled responses are censored so that the rapid, online circulation of the message is less likely to encourage mass protest or breed opportunities for the regime's political enemies to exploit. Recent research by King, Pan, and Roberts (2013, 2017) focused on the Peoples' Republic of China (PRC) has suggested a more nuanced approach to online censorship strategies, one that needs to be balanced between the

CONTACT Elliot Ji esji@princeton.edu Princeton University, USA

[1]Karen Rasler, 'Concessions, Repression, and Political Protest in the Iranian Revolution,' *American Sociological Review*, 1996, 132–52.

[2]Courtenay R Conrad and Emily Hencken Ritter, 'Treaties, tenure, and torture: The Conflicting Domestic Effects of International Law,' *The Journal of Politics* 75, no. 2 (2013): 397–409.

[3]Anthony Oberschall, *Social Conflict and Social Movements* (Prentice-Hall Englewood Cliffs, NJ, 1973); J Craig Jenkins and Charles Perrow, 'Insurgency of the Powerless: Farm Worker Movements (1946–1972),' *American Sociological Review*, 1977, 249–68; Hardin Russell, 'Collective Action,' *RFF Press* 150 (1982): 2000. Conversely, scholars such as Rasler (1996) and Bell, Cingranelli, Murdie and Caglayan (2013) found that violent repression could also lead to more protests and thus increased insecurity.

state's need for control and the private, online service provider's need to stay profitable.[4] Currently, the scholarly community's efforts to investigate online censorship have been confined to the public sphere of the Internet or social media platforms; information that is publicly available and therefore its consumption a result of self-selection by a denizen of the Internet. While this literature has increased our leverage over both the theoretical and empirical implications of online censorship, fewer studies have examined the censorship strategies employed in the semi-private domain or the network level of social media. An arguably more direct route for potentially discordant or subversive information to be delivered as individual users are networked into a group of peers, they both know and likely share the political opinions of.

This article fills this gap by providing experimental evidence that the Chinese government is more sensitive to information exchanged and shared at this network level. What is more, the PRC is not only apt to treat semi-private information differently than public, but also actively attempts to dissuade users from sharing similar information in the future. It is theorized that because of the more intimate nature of the semi-private, social media network, calls to collective action should be more likely to encourage the organization of protests, demonstrations, and similar activities that could be politically disruptive.

This article continues in four parts with a review of the relevant literature and the presentation of the theoretically driven hypotheses. After explaining the experimental design, results and interpretations of the findings follow. Finally, the conclusion includes comments about this and future works on censorship.

Literature Review

The CCP has been institutionally building and enhancing its censorship regime starting with print media in the 1950s.[5] Three major categories of censorship strategies since this time have been identified. Firstly, there is shielding, which encompasses both content blocking and Internet filtering, is a technical method that prevents Internet users from accessing certain, unwanted information on the web.[6] The shielding method is likely to be the most used in the censorship arsenal by authoritarian states. According to Pan 14 authoritarian countries that had Internet penetration of over 40% by 2014 'engage in content blocking related to political and/or religious topics.'[7] The author further suggests that this could be the result of the relatively low technological barrier for implementing Internet filters comparing to other common methods such as content removal. The shielding method is most known for blocking common websites such as Facebook, Twitter, and Google in China. Studies have identified the Great Firewall (GFW) as part of the online surveillance and control system, citing that its 金盾工程 [Golden Shield Project] is the primary censoring apparatus that blocks any information that is deemed prohibited materials by the Chinese government.[8]

The second strategy is what is known as erasing, or the content removal process where information is removed from the Internet and cannot be easily recovered. Erasing involves permanently deleting undesirable information from the forum servers. After removal, those who have not yet been made aware of the information would never be able to learn about it actively (unless being told

[4]Jennifer Pan, 'How Market Dynamics of Domestic and Foreign Social Media Firms Shape Strategies of Internet Censorship,' *Problems of Post-Communism* 64, no. 3–4 (2017): 167–88; Ashley Esarey, *Speak No Evil: Mass Media Control in Contemporary China* (London, Freedom House, 2006); Ashley Esarey, 'Cornering the Market: State Strategies for Controlling China's Commercial Media,' *Asian Perspective* 29, no. 4 (2005): 37–83.

[5]Chin, Sei Jeong. (2018). Institutional Origins of the Media Censorship in China: The Making of the Socialist Censorship System in 1950s Shanghai, *The Journal of Contemporary China*, 27(114): 956–972.

[6]Pan, 'How Market Dynamics of Domestic and Foreign Social Media Firms Shape Strategies of Internet Censorship'; Beina Xu, 'Media Censorship in China, Council on Foreign Relations, 7 April 2015,' *Source Can Be Found*: https://www.Files.Ethz.Ch/Isn/177388/Media%20Censorship%20in%20China.Pdf, n.d., 1.

[7]Pan, 'How Market Dynamics of Domestic and Foreign Social Media Firms Shape Strategies of Internet Censorship.'

[8]Bin Liang and Hong Lu, 'Internet Development, Censorship, and Cyber Crimes in China,' *Journal of Contemporary Criminal Justice* 26, no. 1 (2010): 103–20; Jonathan Zittrain and Benjamin Edelman, 'Internet Filtering in China,' *IEEE Internet Computing* 7, no. 2 (2003): 70–77; William Thatcher Dowell, 'The Internet, Censorship, and China,' *Geo. J. Int'l Aff.* 7 (2006): 111.

by someone who knows), nor can they seek additional information on the matter since the content is deleted forever. The content removal method, also known as the '404' action (the error code for page not found) is believed to be the most effective method to reduce collective action potential and prevent dissemination. There is also strong empirical evidence that content rich in collective action potential are the most likely to be removed from the internet.[9] The removal could be as quick as within a day, while anecdotal evidence suggests that it may only take hours or minutes to remove content from already flagged authors.[10]

This effectively prevents the potential spread of information, and thus undercut the ability for netizens to engage in collective action. Further, content removal conveniently avoids the public's awareness of the censorship establishment, creating an illusion of a safe Internet for users. The Chinese people also have a propensity to support Internet censorship especially among the young.[11] The erasing method, in this sense, also contributes to the establishment and stability of the censorship regime. A perfect censorship system would be one that is perceived as nonexistent by the public. To an extent, swift, surgical removal of unwanted content is the means of achieving this goal.

Unlike the two methods that are discussed above, the unplugging method is a set of legal frameworks that formalize the mandates private, tech companies face to actively censor. This would entail police effort to arrest, detain, and silence unwanted info sources; or legal means to ban the release of information before it was made known to average Internet users. In other words, the government unplugs the source from which undesired information originates in the real world. In 2007, the state-owned Internet Service Provider (ISP) Telecom in Henan Province shut down four hundred servers for noncompliance with the censorship laws and regulations. In the same year, two local ISPs in Chengdu, Sichuan Province, 世纪东方 [Century Oriental] and中客科技 [Zhongke Technology] were denied access to the Internet for allowing prohibited materials on their websites.

As for silencing individuals who author the 'prohibited materials and information', the Chinese government also removes them from cyberspace by physically persecuting and then incarcerating them under various charges, most common of which being 'inciting subversion of state power.' According to *Reporters sans Frontieres*, 60 Chinese nationals have been constrained (house arrest), detained, arrested, and/or persecuted for authoring or spreading dissenting information.[12] Among the list of 69 figures, there is prominent CCP dissenters such as Liu Xiaobo, former CCP officials, academics such as Li Jianping, lawyers, reporters, civil rights activists, even average netizens who commented on popular political issues.

The impact of the unplugging method on both public and the non-public domain is that it breaks some critical connection points in the dissenting network that are crucial to expanding the network and communicating among activist groups.[13] The unplugging method could be especially detrimental to dissenting networks in the private domain such as WeChat users. WeChat utilizes a 'point-to-point' connection model that the user can only initiate communication with someone who approves the friend request. Removing one core user in a WeChat network could render a group of users isolated from another group in which the

[9]Gary King, Jennifer Pan, and Margaret E Roberts, 'How Censorship in China Allows Government Criticism but Silences Collective Expression,' *American Political Science Review* 107, no. 2 (2013): 326–43; Gary King, Jennifer Pan, and Margaret E Roberts, 'Reverse-Engineering Censorship in China: Randomized Experimentation and Participant Observation,' *Science* 345, no. 6199 (2014): 1,251,722.

[10]Tao Zhu, David Phipps, Adam Pridgen, Jedidiah Crandall, and Dan Wallach, 'The Velocity of Censorship: High-Fidelity Detection of Microblog Post Deletions,' in *Presented as Part of the 22nd {USENIX} Security Symposium ({USENIX} Security 13)*, 2013, 227–40.

[11]Steve Guo and Guangchao Feng, 'Understanding Support for Internet Censorship in China: An Elaboration of the Theory of Reasoned Action,' *Journal of Chinese Political Science* 17, no. 1 (2012): 33–52.

[12]'RSF's 2018 Round-up of Deadly Attacks and Abuses against Journalists—Figures up in All Categories,' *Reporters Without Borders* (blog), 18 December 2018, https://rsf.org/en/news/rsfs-2018-round-deadly-attacks-and-abuses-against-journalists-figures-all-categories.

[13]Alex Rutherford, Manuel Cebrian, Sohan Dsouza, Esteban Moro, Alex Pentland, and Lyad Rahwan, 'Limits of Social Mobilization,' *Proceedings of the National Academy of Sciences* 110, no. 16 (2013): 6281–86; Elisabeth R Gerber, Adam Douglas Henry, and Mark Lubell, 'Political Homophily and Collaboration in Regional Planning Networks,' *American Journal of Political Science* 57, no. 3 (2013): 598–610.

core user is the only mutual contact with the first group. As a result, re-grouping or re-organizing in the private domain may prove much more difficult with the unplugging method in practice.

Previous literature has established that authoritarian regimes impose controls over media sources and forums for public discourse, and many studies have investigated the strategies, motivations, and logistics of the censorship regime in China.[14] Prior to the rise of the Internet and social media sites, state-sponsored censorship had been practiced with traditional media, such as newspaper and radio broadcast.[15] State-owned media outlets enable the state to selectively disseminate information and thereby remove the need to apply the process of censorship. In contrast, there are various influences a state could exert on private media outlets which could coerce these outlets into submission. This can be done by forcing private media outlets to accept state scrutiny of content before its release, allowing state agents to oversee broadcasts, or making the consequences of violation uncertain in order to promote self-policing among the private media platforms.[16]

However, with the rise of the Internet and social media, a new challenge for states has emerged in that information is disseminated (and sometimes distorted) at great speed, leaving the state's ability to precisely censor materials weakened.[17] As a result, states have turned to new technologies to impose a regime of broad-spectrum censorship on online discourse to censor-worthy information and quickly.[18] The effectiveness of this approach has been viewed positively by several scholars, finding that states can employ effective censorship through collaborating with or infiltrating Internet service providers.[19] China, being one of the most studied authoritarian regimes today, is often regarded as a success story in state efforts of information control.

In recent years, with the rise of smartphone-based instant messaging apps, the State Council authorized the Cyberspace Administration of China to create provisions that solely target what is called the '公众号[Official Accounts]' These are private news or opinion outlets most used on WeChat. In fact, four out of eleven bills and provisions passed and released on the matter of Internet control in 2017 aimed at monitoring and censoring online forums in both public and semi-private platforms. With potentially more regulations and legislation on the way, the internet censorship regime has quickly grown into a mature, efficient state apparatus that can reduce, control, and eliminate unwanted information from cyberspace.

Scholars have also investigated the censorship regime in China from an accessibility angle, arguing that the Great Firewall, a cyberspace gatekeeper blocking the Chinese citizens from accessing certain websites, is an attempt to reduce the likelihood of collective action triggered by foreign

[14]Pippa Norris and Ronald Inglehart, *Cosmopolitan Communications: Cultural Diversity in a Globalized World* (Cambridge, UK: Cambridge University Press, 2009).

[15]Geoffry Taubman, 'A Not-so World Wide Web: The Internet, China, and the Challenges to Nondemocratic Rule,' *Political Communication* 15, no. 2 (1998): 255–72; Pan, 'How Market Dynamics of Domestic and Foreign Social Media Firms Shape Strategies of Internet Censorship.'

[16]Jonathan Hassid, 'Controlling the Chinese Media: An Uncertain Business,' *Asian Survey* 48, no. 3 (2008): 414–30.

[17]Peter Ferdinand, *The Internet, Democracy and Democratization* (Routledge, 2013); Philip N Howard, Aiden Duffy, Deen Freelon, Muzammil M Hussain, Will Mari, and Marwa Maziad., 'Opening Closed Regimes: What Was the Role of Social Media during the Arab Spring?,' *Available at SSRN 2595096*, 2011; Gilad Lota, Erhardt Graeff, Mike Ananny, Devin Gaffney, and Ian Pearce, 'The Arab Spring| the Revolutions Were Tweeted: Information Flows during the 2011 Tunisian and Egyptian Revolutions,' *International Journal of Communication* 5 (2011): 31.

[18]King, Pan, and Roberts, 'How Censorship in China Allows Government Criticism but Silences Collective Expression'; King, Pan, and Roberts, 'Reverse-Engineering Censorship in China: Randomized Experimentation and Participant Observation'; Peter Lorentzen, 'Regularizing Rioting: Permitting Protest in an Authoritarian Regime,' *Social Science Research Network* 995,330 (2010): 1–45.

[19]Evgeny Morozov, *The Net Delusion: The Dark Side of Internet Freedom* (Public Affairs, 2012); Rebecca MacKinnon, 'Flatter World and Thicker Walls? Blogs, Censorship and Civic Discourse in China,' *Public Choice* 134, no. 1–2 (2008): 31–46.

actors and liberal ideologies.[20] Others, in contrast, looked at online censorship on social media websites, expanding the argument of collective action prevention to a more intimately connected environment.[21] Their works suggest that the Chinese government strategically removes content on social media sites and blocks the internet provider (IP) addresses of opinion leaders intending to stop their collectivizing before observable action can occur. This argument is empirically backed by large-scale studies that collected and analyzed millions of censored social media posts.[22] The studies found that the censorship mechanism tends to target posts with a high collective action potential and spare those that are merely critical of the state, and by denying access and throttling internet speed in the public domain, the government not only undermines the collectivization capacity of the population but also strategically distracts the attention of the public opinion. The central theme remains the regime's concern for large-scale collective action.

However, the extant literature has yet to address censorship performed online within closed, social networks and other semi-public domains. They focused mostly on the censorship imposed on public forums and information that is accessible to any internet user. At this level, it is expected that their potential for creating opportunities for collective action is much weaker. The platforms that operate on a more private, close-knit level, in contrast, differ considerably in terms of their capacity to organize collective action. Dovetailing on the discussion of collective action prevention as the objective of state censorship, this article aims to provide an account of the cross-domain variation of censorship strategy that expands beyond the public domain.

Theoretical Development

This article departs from the existing works by noting a critical theoretical distinction between the two domains in terms of their collectivizing capacities. It is argued that the semi-private domain has a unique network structure among socially homophilic users that better collectivizes these actors in real life. Namely, the semi-private social media platforms encourage users to bond and see communities through acquaintances, creating obstacles for total strangers to join a group without having known someone in that group or sharing something in common with the group members. This will likely enable the semi-private platforms to better disseminate coordinating information among like-minded people and reduce concerns of unreliable information, boosting their ability to breed successful collective action. Then, applying the empirically substantiated theoretical claims of existing literature, the censorship strategy deployed on semi-private platforms needs to be different than that used on the public platforms to achieve the same level of efficiency in stifling collective actions.

The Public Domain and Collective Action

The collective action problem is one of political science's most oft-discussed topics.[23] At the street level, citizens do not always share the same preferences nor respond to the same incentives, meaning the 'problem' is encouraging and rewarding collective action from an unorganized

[20]Margaret E Roberts, *Censored: Distraction and Diversion inside China's Great Firewall* (Princeton University Press, 2018); Zittrain and Edelman, 'Internet Filtering in China'; Jack Linchuan Qiu, 'Virtual Censorship in China: Keeping the Gate between the Cyberspaces,' *International Journal of Communications Law and Policy* 4, no. 1 (1999): 25.

[21]Pan, 'How Market Dynamics of Domestic and Foreign Social Media Firms Shape Strategies of Internet Censorship'; Guo and Feng, 'Understanding Support for Internet Censorship in China: An Elaboration of the Theory of Reasoned Action'; Liang and Lu, 'Internet Development, Censorship, and Cyber Crimes in China.'

[22]King, Pan, and Roberts, 'How Censorship in China Allows Government Criticism but Silences Collective Expression'; Roberts, *Censored: Distraction and Diversion inside China's Great Firewall*.

[23]Elinor Ostrom, 'A Behavioral Approach to the Rational Choice Theory of Collective Action: Presidential Address, American Political Science Association,' *American Political Science Review* 92, no. 1 (1998): 1–22.

mass.[24] The 'solution' to this problem for many groups can result in political mobility, and often, from the state's point of view, an increased threat to the regime's stability. Thus, collective action is something that regimes might be tempted to dissuade to preserve their security.

Public online platforms provide a pool that gathers, retains, and spreads collectivizing information to its readers. At the minimum, the readers must be able to 1) reliably receive the message, and 2) consider this message credible so that they can coordinate their actions based on the message and collectivize. This holds true for even the most adamant dissenters since it lays the logistic foundation of the successful organization of participants. In other words, for a platform to have a high potential of breeding collective actions, it must be able to consistently meet the above two criteria.

But are public domain platforms meeting those basic requirements? Existing literature that has presented large-N quantitative empirical work seems to doubt it, especially concerning the credibility of messages spreading in public groups. Recent pieces on censorship have demonstrated that public social media platforms tend to form network clusters that are more likely to share information among nodes with established connections. This allows the government to capture the population's online discourse by topic groups.[25] By its nature of being publicly accessible, the public domain often requires little to no identity verification process before users can access and post content. Note that this verification process does not need to be in the form of a real-name registration. In private, point-to-point messaging platforms, this verification is as simple as a friend request; whereas in public platforms like 知乎 [Zhihu], the Chinese version of Quora, users have no way of identifying who the other user is. This burdens each user who is potentially considering joining a collection with the task of differentiating the 'true calls' from 'rumors'. This will likely exacerbate the collective action problem; the natural suspicion among users renders large-scale collective action difficult to stem from calling-for-action posts in public platforms like Weibo.

The Semi-Private Domain

The semi-private, social media network has arguably overcome the 'collective' aspect of this problem. The semi-private domain in this study is defined as the platform on which information is passed and disseminated but requires certain membership or subscription procedures that limit the audience to a group of people with some connections rather than the general body of 'netizens'. The reason it is semi-private is that the information being sent through a semi-private platform is not completely private (i.e. only visible to the sender and the receiver) or completely public (i.e. visible to anyone on the internet). There is either an element of subscription or a requirement of membership to receive the information passed by the sender. In other words, the members have already been collected based on some number of shared interests or common traits. This could be wholly benign, such as a group dedicated to pop culture fandom, or, and to the state's chagrin, a group with shared antipathy to the regime.

In either case, these members' ability to both seek out like-minded peers and then network with them was facilitated by the semi-private social media platform. Social networks, long before the advent of the Internet, were an important part of collective action studies with key differences in findings usually falling on the size and density of the network.[26] Messages circulating among friend groups and acquaintances are more likely to be found trustworthy by the receivers, rendering them less hesitant when committing to collective action. Furthermore, an additional mechanism may be

[24]Mancur Olson, *The Logic of Collective Action: Public Goods and the Theory of Groups*, Second (Cambridge, MA: Harvard University Press, 2009).

[25]Jennifer Pan and Kaiping Chen, 'Concealing Corruption: How Chinese Officials Distort Upward Reporting of Online Grievances,' *American Political Science Review* 112, no. 3 (2018): 602–20.

[26]John T Scholz, Ramiro Berardo, and Brad Kile, 'Do Networks Solve Collective Action Problems? Credibility, Search, and Collaboration,' *The Journal of Politics* 70, no. 2 (2008): 393–406; David A Siegel, 'Social Networks and Collective Action,' *American Journal of Political Science* 53, no. 1 (2009): 122–38.

that since messages are being sent through a subscription process, they can be targeted to the preferences of the like-minded audience, achieving a more collectivizing effect than 'mass email' type callings commonly found on public domain platforms. In short, smaller, more centralized groups are more consistently found to overcome the collective action problem.[27] These groups are better able to organize their efforts and maintain similar interests and incentive structures.

The network of WeChat users is a semi-private network where information can only flow between users who have established a direct connection (i.e. being 'friends' on the app). Posts on this platform will only be visible to those with whom you have zero-degrees of separation and authors can only publish to their subscribers. The WeChat platform can then be viewed as creating tighter social connections between acquaintances and peer groups which a user self-selects into. The collective action potential on such a platform should be much higher, meaning that one should expect if not wholly different, certainly conditioned, strategies in relation to censorship. Therefore, existing empirical explorations on this topic must be expanded to those domains beyond the public if one is to offer a comprehensive picture of why Chinese internet censorship is effective.

It is argued that censorship strategies employed on the semi-private domain are more indiscriminate than those on the public domain to counter its higher collectivization capacity. This study aims at testing the mechanism of censorship in the semi-private domain using an experimental research design. WeChat, the single most popular instant messaging system in China, was employed as it has global user groups and can reflect the flows of information in both the private and the semi-private domains. Operationalizing the difference in collective action potential between the public and the semi-private domain, the hypothesis below to reflects the pattern of censorship in the semi-private domain. Because of the nature of the semi-private domain, the state is expected to be just as sensitive to posts of this nature as they would those calling for action. The central hypothesis of the paper is thus written to reflect the expected indiscriminate pattern of censorship in the semi-private domain.

Hypothesis: State critique posts on semi-private social media platforms are no more likely to be censored than their calling-for-action counterparts.

Research Design

Based on the existing literature on censorship theories, this study categorizes the censor-worthy materials into the following groups: political criticism (or dissenting opinions), call-for-action, historical sensitive topics, and current sensitive topics. The political criticism group includes materials that are considered dissenting opinion or words that question the party's or the leader's legitimacy, performance, validity of policy and decisions, the rule of law, etc. This group contains information that has high resonance among educated citizens and can be translated into mass mobilization among the public. It is in the regime's interest to preserve the veneer of stability, prosperity, and high performance of the government. Therefore, such political criticism falls under the radar of censorship programs. The call-for-action group is, by its name, referring information that aims at organizing collective actions (e.g. protest, sit-ins). This is particularly notable among student groups or civil societies, where mobilization among like-minded people faces fewer barriers compared to those among strangers.

Additionally, sensitive topics, both historic and current, need to be discussed separately. The former group collects discussion of history-defining events such as the 1950 Invasion of Tibet, the Great Leap Forward (1958–1962) and the Great Chinese Famine during that time, the Cultural Revolution (1966–1976), the Tiananmen Square Incident (1989), several self-immolation incidents

[27]Gerald Marwell, Pamela E. Oliver, and Ralph Prahl, 'Social Networks and Collective Action: A Theory of Critical Mass,' *Journal of Sociology* 94, no. 2 (1988): 502–534.

(related to Tibetan issues and the Falun Gong), *etc*. Historical text on these events is usually heavily redacted and altered, if not completely disappeared. Politically controversial events that occurred as breaking news after 1989 without long-term impact would be put into the group of topics that are currently sensitive. This group would, for instance, include political speculations that arise after the arrest of a high-profile party official, such as Bo Xilai, Xu Caihou, or Zhou Yongkang. The members of this fourth group typically have a very short half-life in terms of salience, usually fading from the public's attention within weeks.

The first phase of data collection focuses on using posts that are critical of the government and those that directly call for action. This categorization of content is like that used in a previous study which conducted a large-scale randomized experiment to discover the precise mechanism by which Chinese censorship operates on sites that are publicly available to Internet users such as social media sites and discussion forums (e.g.天涯论坛[Tianya Forum] and腾讯微博[Tencent Weibo]).[28] Their experiment was able to create many posts and monitor their status in various locations, but the scope of their study is limited to the public domain while the flow of information is also active in platforms that require membership or subscription to access. Here, the empirical context is expanded to the semi-private domain with a more clear-cut categorization scheme of the post content to enhance and supplement extant findings in the literature.

The semi-private domain in this study is operationalized as a platform on which information is passed and disseminated but requires certain membership or subscription procedures that limit the audience to a group of people with some connections rather than the general body of internet users. An example of such a platform is achieved via the official account on WeChat which is the platform on which to conduct the experiment. The official accounts are essentially publishers that can send posts or messages to their subscribers. The posts can be 'shared' like Facebook posts and articles by the subscribers to their contacts (See Appendix A for illustration). The content of the post is essentially a link registered under the WeChat platform. Therefore, when one post is censored, the link is re-directed to a page informing the reader that the article is no longer available. If a post is not censored, the link remains valid. Any user who has the link can access the content of the post if the official account that authored the post is renewed and maintained properly. However, a regular official account can only push one article each day.[29]

The authors created a dummy official account from which to issue the experiment's posts. This account's name was intentionally crafted to be politically benign to avoid its being flagged by the system.[30] If the King, Pan, Roberts hypothesis holds, it should be expected that the posts that are high in collective action potential will be quickly censored while the posts that are merely critical of the regime will be spared. Therefore, the collective action group posts are created with a clear call-for -action subject. To ensure that these posts can reasonably approximate the readability of posts authored by real users, all experimental posts were 1300–1500 characters in length, the average length of the 50 uncensored posts the researchers collected that have more than 200 reads each. Additionally, as mentioned above, the dummy official account will have 30 dummy subscribers. This allows the posts that are published via the dummy official account to all have the same number of reads (30) and thus eliminate post popularity as a potential confounder to the study. To measure the survival time for each post, once the posts are published on the official account with a valid link, the webpage will be automatically refreshed in five-minute intervals. The 'refresher' is pre-programmed to stop as soon as the browser is re-directed to a new page, suggesting that the original content has been censored. Once the refresher stops, the number of times it refreshed the page is recorded and how long the post survived calculated. Since the refresher is set to refresh at five-minute intervals, the margin of error in determining how efficient the censorship is meaningfully negligible.

[28]King, Pan, and Roberts, 'Reverse-Engineering Censorship in China: Randomized Experimentation and Participant Observation.'
[29]微信公众平台服务协议[*WeChat Official Account User Agreement*], 2017.
[30]The name of this account has been omitted here at the suggestion of Emory University's IRB.

A baseline was established for the WeChat censorship system using a series of control group posts. These included posts that explicitly refer to historically sensitive materials (which should be censored rather quickly), as well as posts that are no more than gibberish. Using a small corpus of posts the authors were able to salvage before they were censored, content analysis was used to identify the 'censor-worthy' key phrases such as '造反[to revolt]' and '独裁[dictatorship].'

As an exercise to reveal algorithmic censorship, a test to determine the extent of human censors' involvement in WeChat censorship is performed. While preserving the same number of key phrases, the text of a post is scrambled. The scrambled text is evident to a human eye, as the sentences make no sense. However, for a computer algorithm programmed to flag keywords, the post will be just as censor worthy as they were before the treatment. For example, the post:

'…the Constitutional Amendment that granted life tenure to the president is a serious institutional setback that directly contradicts with Deng Xiaoping's reforming ideas' (主席任期终身制是严重的制度倒退, 与邓小平的改革思想背道而驰……)

which abstractly criticizes Xi Jinping's attempt to grant himself tenure for life was turned into the following:

'…directly/to/that/serious/the/that/with/Deng/ideas/Xiaoping's/reforming/Amendment/tenure/a/life/institutional/contradicts/setback/president/Constitutional/granted/is/the …'

In general, if human censors are examining the posts after they have been flagged by the system as a form of 'quality control,' one could expect that these scrambled texts would be allowed on the WeChat platform since whatever information they could convey is now indecipherable to human readers. However, if the censorship is conducted mostly by an automated algorithm, these scrambled posts would not have much chance escaping scrutiny as their original texts were all previously flagged and censored. To establish a baseline for comparison between the original and the scrambled text survival, all original text, though known to have been censored previously, were re-posted through the dummy official account and have their survival rate and time measured.

For the main group of comparison, two groups of posts are created reflecting content that is critical of the state and calling for action, respectively. The collective action group posts are directly calling for action on matters that are concerning to a group of citizens (in this case, secondary education). An *a priori* guideline is employed for selecting and creating this group of posts to ensure a valid comparison with the treated group in terms of the content. To be categorized as 'calling-for-action', the post must 1) address an identifiable sub-group of the Chinese population; 2) have a specific time and location for organized action; 3) address a specific grievance that does not apply to at least one other sub-group (e.g. unfair examination policy, exorbitant tax burden, etc.). The last guideline is a conservative approximation of the nature of most organized protests in China: self-contained and small-scale with a grievance that is usually localized. Below is an example of what could constitute a collective action post:

"…have dozens of our people…petition to the Commission of Education, sit-in. Make our voice heard!" (几十人……去教委情愿, 静坐让他们听到我们的声音!)

The state critique group comprises of posts that are also authored by the researchers and are within the character limits as the collective action group posts are. The only difference is that instead of calling for real-life actions such as protests, the treated posts are written to reflect academic criticisms of the Chinese government, domestic policies, and general political debates abstractly. In other words, the treated posts resemble more of an op-ed with an argument critical to the state than they do of a poster for propaganda purposes. These posts are written or revised from real-world posts based on the following *a priori* guideline: 1) The posts will not include any locations, time, or their intended audience but focus exclusively on the argument that is criticizing the policy, the leader, and the regime; 2) the post will have a direct reference to either a CCP leader or an executive official above the provincial level; 3) the post will explicitly identify a policy change or ideological approach to governance. The reason for adopting this conservative *a priori* guideline was to provide methodological assurance against cherry-picking posts and assign them to categories ex-post based

on their time to reduction. The study can thus clearly differentiate between posts that are only criticizing the state with little intention to organize collective action and the posts are calling for actions.

Like the calling-for-action group, the state critique group posts will also be refreshed in a five-minute interval and be measured in the same manner. The King, Pan, and Roberts findings suggest that the proportion of posts that are censored and the survival time of state critique posts to be significantly different from those of the calling-for-action posts. Namely, the proportion censored would be much less while the survival time would be much longer.

To protect data integrity and prevent real-life political implications to private citizens, this study only shares information with the 30 WeChat personal accounts created by the researchers, from the one, dummy official account. Per the微信公众平台服务协议[WeChat Official Account Service Agreement], personal official accounts that are not verified by the platform (i.e. having no established link to companies or institutions such as schools) are only permitted to publish one post a day. The post is immediately published to its subscribers with a unique link to the readers. Following the same protocol, the automatic webpage refresher is set to refresh in a five-minute interval. Posts' survival times are recorded with a 5-minute margin. To keep the analysis conservative, the lower bound of a post's survival time is recorded. If a post remains accessible 48 hours after being published, it is classified as having 'survived.'

Results and Analysis

The experiment consists of 256 total posts across the above four 'types:' historic, gibberish, collective-action, and state critique. Table 1 summarizes survival time for the censored posts by post type. Consistent with the general expectation, the control posts (i.e. the content that had been previously censored) are removed quickly whereas the same content, upon being scrambled took much longer to be censored.

Why might algorithmic censorship be slower than human censorship? The first conjecture, based on some informal conversations with software engineers and content reviewers at popular social media platforms in China, is that algorithmic censorship still must rely on keyword detection, but it probably employs a more sophisticated keyword-in-context analysis to look for certain syntax or conjugations to ameliorate the high false positives limited keyword-based methods have. This would explain why the scrambled texts do survive longer despite being processed from posts that had been censored before. Once turned into gibberish, the keywords fall out of context, forcing such an algorithm either to clear it or pass it to human review. This may suggest that while an algorithm is certainly in place to do an initial pass of the content, the platform is keenly aware of its limit. The platform has an economic incentive to avoid false positives because excessively censoring the posts can discourage users from using its social media service. Hence, not only are platforms incentivized to maintain minimum compliance with the state's censorship practice, but they are also sensitive to false positives such that they are willing to employ many human reviewers to process the flagged posts.

However, as the results for the historic posts suggest, the algorithm may be allowed to remove posts that are censored without human review, meaning that human reviewers are not the only actors of the censorship regime. This is the second conjecture. In these trials, when seeing sensitive

Table 1. Summary of survival statistics by post type.

Type	Observation	Mean	Stnd. Dev.	Min	Max
Scrambled	57	174.02	308.36	0	1205
State Critique	56	24.37	61.16	0	335
Collective Action	54	16.60	35.85	0	165
Historic	59	5.54	12.18	0	70

Observations that are marked 'survived' are not included in this table.

content that has been previously censored, the algorithm removes posts almost immediately, allowing little to no time for human review and decision-making. Intuitively, there would be no point to have the human reviewers examine posts that were repeating a previously flagged article verbatim. Strategically, having the machine weed out reoccurring undesirable content is valuable during volume bursts such as when an event goes 'viral.' Official accounts tend to copy and re-post the same article (which would also explain why some content can go viral rapidly). The results from the gibberish posts indicate that there might be something more to the existing understanding on algorithmic censorship use although nothing conclusive can be offered.

Overall, these findings partially support existing findings in the sense that they show humans are likely involved in deciding to remove a post in different domains. This finding should be considered a supplement to the King, Pan, and Roberts findings because this provides evidence that human reviewers are involved to a great extent, but algorithms are also more than content sniffers—they destroy as well as detect.

Figure 1 demonstrates the differences in survival time between the latter two types. Both groups of the post have a noticeable decrease in survival time as the 13th National People's Congress (NPC) approaches. The NPC lasts for 16 days from March 5th to March 20th of 2018. Both groups also show a rebound in survival time after the meetings conclude. The two groups have similar patterns of change in survival time before, during, and after the NPC with plummeted survival time during the meetings. Overall, there is little discernible difference in surviving WeChat censorship between the two kinds of posts.

But are these findings significant? A Cox proportional hazard model is used to determine if there is a difference between one type of post from another. Here, instead of a single post being the unit of analysis, it is post-time with each observation maxing out at 100 seconds. In this way, a single post might create multiple observations, and the N increases from 256 to 657. Again, the suspicion is that gibberish posts should be the most likely to survive, and those posts with historically sensitive

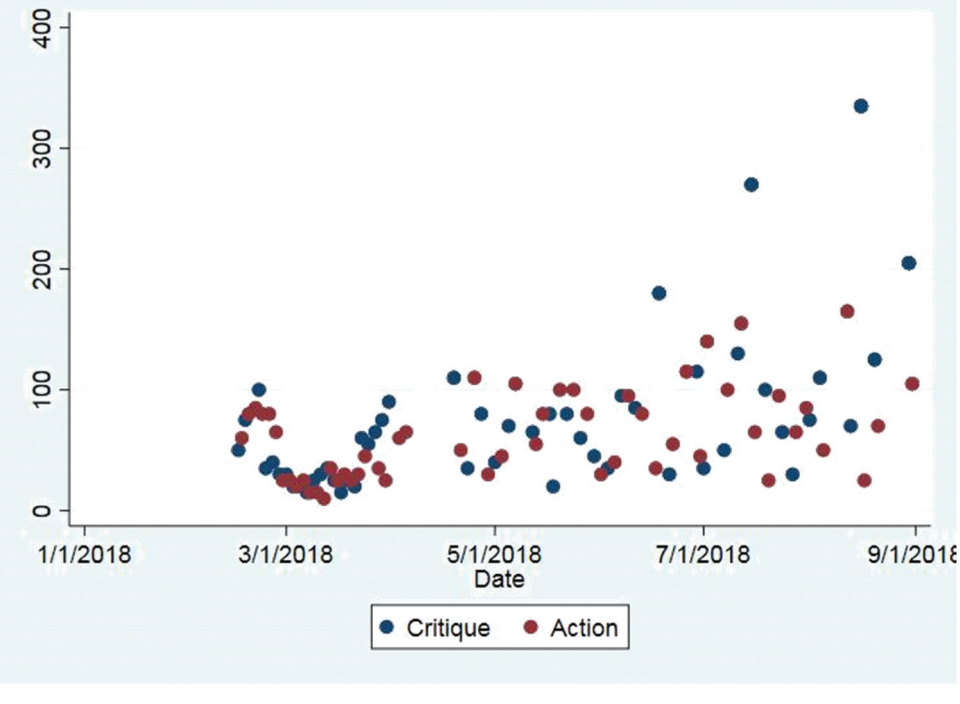

Figure 1. Difference in survival for critique and collective action posts.

content the most likely to be redacted. Figure 2 below demonstrates the difference in a post's likelihood of being censored dependent on its contents. While a state critique post is more likely to survive than a collective action post, both are certainly more likely to be censored than the gibberish posts.

The hazard ratio for this model, reported in Table 2 below, is somewhat striking; it suggests that as a post moves from gibberish to historic content, it is 3.5 times more likely to be censored.

Concerning the central hypothesis, no significant evidence that there are any statistically discernible differences between the survival rate of a post that is specifically calling for collective action and that of an abstract critique of the Chinese government or the CCP (Table 3) was found. The posts that have critical content against the state are surviving on average slightly longer than their calling-for-action counterparts, but the difference is barely significant at the 0.05 level. Table 4 reports a power analysis to determine the necessary N for this comparison of hazard ratios to achieve the standard power of 0.80. Only 64 observations are required to achieve the standard power. This

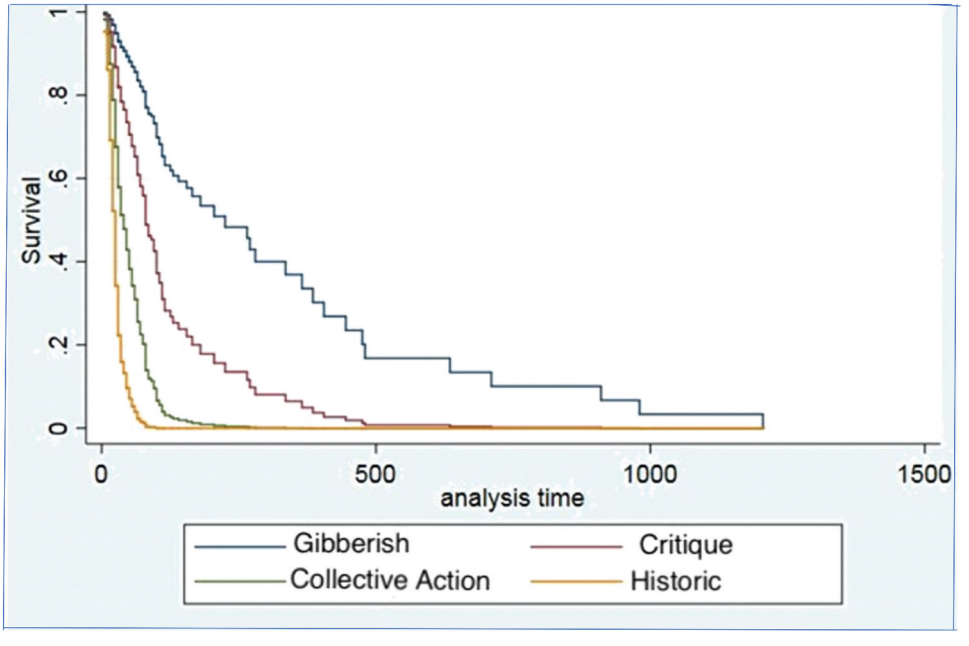

Figure 2. Cox proportional hazard regression.

Table 2. Hazard ratio & coefficient, state critique vs. collective action.

	Hazard Ratio	Coefficient	95% Conf. Int.
From Gibberish to Historic Posts	3.515	1.257***	2.291–5.393
		(5.76)	
N			256

t statistics in parentheses; *** p > 0.001

Table 3. Hazard ratio & coefficient, control vs. treated posts.

	Hazard Ratio	Coefficient	95% Conf. Int.
Critique to Action Posts	1.948	0.667	0.683–5.556
			(1.25)
N			84

t statistics in parentheses; *** p > 0.001

Table 4. Power analysis for all pairs of comparison.

Variable	Hazard	H1—H2/3	N	Power
Gibberish	0.014	-	88	-
Action	1.740	1.726	88	1.0
Critique	1.410	1.396	88	1.0

Alpha: 0.050 (two-tail)

suggests that the WeChat platform has a pattern of censorship that does not conform to the expectations of existing literature and adds empirical strength to our central hypothesis that there is little discernible difference between the post types that King, Pan, and Roberts tested when being switched from the public to semi-private domain.

In sum, the researchers found that the censorship mechanism employed on the semi-private WeChat platform adopts a broad-spectrum filtering method that not only targets posts that are calling for collective action but also those that are merely critical to the state and its policies, very unlike the previous findings on the public domain. The authors attribute this cross-domain variation to the different capacity each domain holds in terms of fostering collective action that is meaningful and significant enough in the eye of the regime to merit censorship. Unlike online forums and discussion pages that are open to any internet user to access and participate, the WeChat official account subscription forms a smaller social network that connects people who self-select to subscribe to the account and thus more exclusive to the general user pool. Therefore, when a message is sent to the subscribers, the owner of the official account can reasonably expect a higher response rate and the likelihood of having real-world gatherings than the public forum users who are broadcasting to total strangers. With the assistance of additional software features that allow the official account owner to collect and disburse information among its subscribers, the WeChat semi-private domain is much more capable of engaging in collective actions than publicly accessible online forums. As a result, the state has incentives to censor more heavily on the semi-private domain due to its greater capability to foster collective action.

Conclusion

The empirical evidence suggests that Chinese political censorship is not simply a one-size-fits-all system of information control but a versatile, highly adaptable instrument that can be customized for a different level of information control. While the exact method of censorship remains mostly constant (content blocking and removal), what merits censorship in the eye of the government is a function of its collective action potential. However, an important distinction is drawn between the environments in which a post with a given collective action potential is disseminated to factor the collective action capacity into consideration. The authors concur with existing literature that the Chinese political censorship has been geared largely toward the prevention of collective action but add to the discussion by suggesting that the censorship strategy varies across domains with different collective action capacity.

It also appears that the level of censorship is more rigorous in the semi-private domain where the degree of separation between users with similar interests is drastically reduced by the design of the official account. Posts that call for immediate collective action are censored as expected while posts that are critical to the regime are similarly censored, unlike other studies have predicted. Furthermore, this article offers concrete evidence suggesting the involvement of human reviewers in the censorship regime. WeChat censors are indeed sensitive to the syntax and internal logic of the posts rather than only flagging keywords and phrases, suggesting that it is very unlikely that screening algorithms alone can identify whether the keywords are in the correct grammatical context (and censor the post) and spare those that are not. Future studies would benefit from expanding the magnitude of the experiment and use multiple official accounts with more dummy

viewers to eliminate the possibility that the rigor of censorship being a function of reader amount. The same method can be further applied and improved in different service platforms to check for robustness. The QQ platform, for example, has a slightly different user pool but is of comparable popularity and function with WeChat.

Further, future studies could also investigate the efficiency of China's censorship regime by looking at its policy roots. How do the Chinese media and cyberspace law and regulations nourish the Chinese censorship regime to be as expansive and as effective as it is today? For instance, one of the key components of Chinese censorship is the government's ability to directly interfere with their service provisions. Regulatory provisions and laws that concern censorship (information control) fall under the larger category of internet regulation, which includes both censorship and other standards that regulate internet services. While the backbone of the legal framework of censorship is built by state legislation (e.g. laws passed by the National People's Congress, administrative regulations by the State Council, local regulations by the local people's congresses, and rules by the central government and local governments) and sometimes by party documents (e.g. policy documents and intra-party regulations), administrative policies and regulations are also active in outlining specific criteria for the censorship regime. The legislative documents, like 全国人民代表大会常务委员会关于维护互联网安全的决定[The Standing Committee of the National People's Congress' Decision on Preserving Cyber Security] which is a political document from the Standing Committee of the NPC in 2000, are often vague, including terms such as '有害信息[harmful information]' to allow the government high discretion when practicing content filtering. The regulations and decisions from agencies, thanks to their dual-capacity in both the state and the Party, can take effect much more promptly without going through a mandatory notice-and-comment and can take effect immediately. To offer a more fine-grained understanding of the use of censorship as an institutional instrument to repress, future studies should go deeper into the legal-administrative source of online censorship through a qualitative angle.

Lastly, with new technology emerging, studies on censorship must be keenly aware of the technological developments that may add to the existing arsenal of censorship and improve research methods accordingly to better account for time inconsistency and change in censorship strategy. Does a strong censorship regime entail strong state capacity or reveals the regime's fear of or its weak spot for civil unrest? As emerging technologies such as artificial intelligence adds to the regime's repressive toolbox, discovering state intent in the planning and implementation of subtle, non-violent repression is vital to promote a better theoretical understanding of repression for the community.

Acknowledgments

The authors thank the panelists of the Digital Technologies and Human Rights Panel at the 2019 International Studies Associations Annual Conference in Toronto, Canada, and two anonymous reviewers and the editorial staff at JCC for valuable comments and suggestions. All errors are our own.

Disclosure Statement

No potential conflict of interest was reported by the author(s).

Notes on contributors

Elliot Ji is graduate student in the Department of Politics, Princeton University

Zack Bowersox is a Lecturer in the Department of Political Science, Washington University in St. Louis Email is bowersox@wustl.edu.

Appendix A

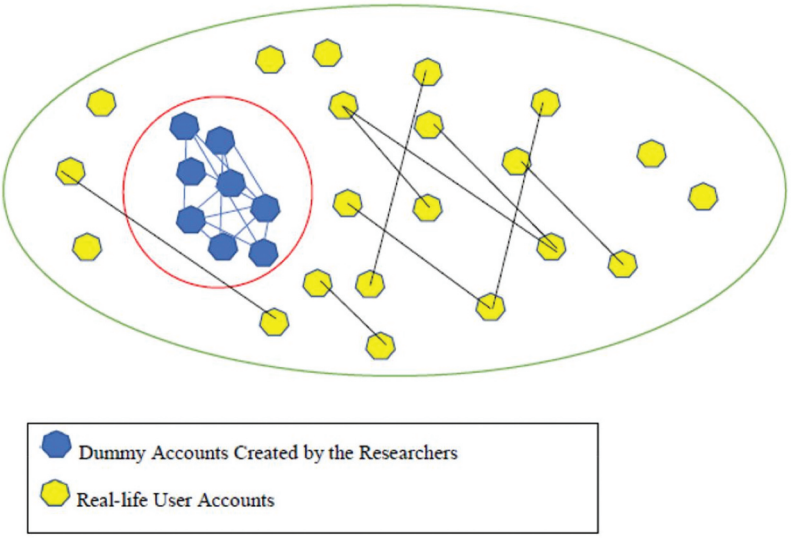

Figure A1. Illustration of dummy account network in the experiment.

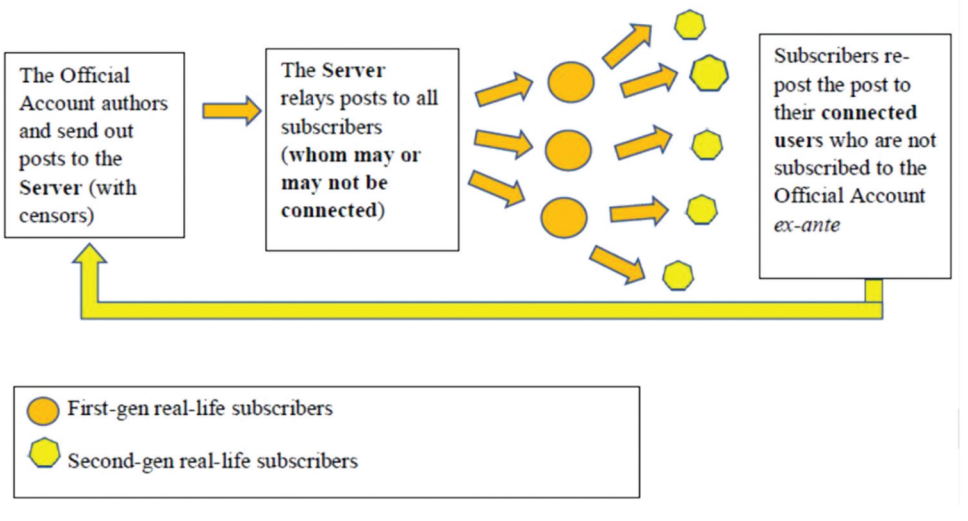

Figure A2. Illustration of real-life official account network.

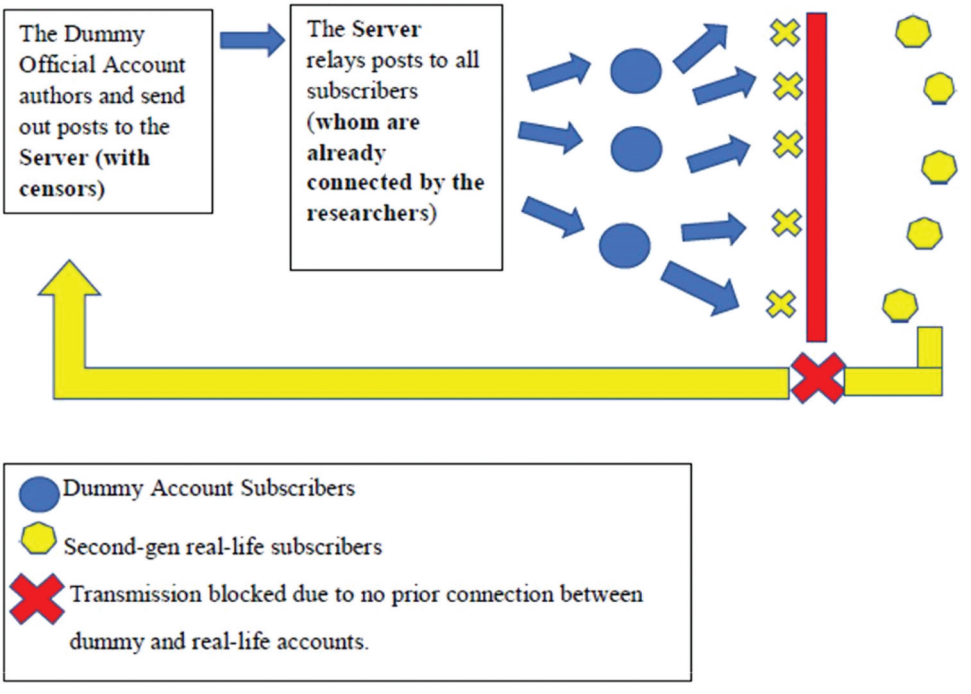

Figure A3. Illustration of experiment official account network (Dummies).

Appendix B

The authors tested posts that fall under the four categories below along with a brief example of each.

(1) Calling-For-Actions: Contents that are calling for specific collective action with explicit purpose, target, and means of protest. The grievance must be applicable to an identifiable sub-group of the general Chinese population such as ethnic minority, prospec-tive college students in a specific providence or region, or those with a shared household registration.

'江苏省考生凭什么要比北京考生多考数十分才能上一样的大学?这是教委的政策歧视!为了我们子女的未来, 我们要发动群众, 组织数十人去南京教委门前静坐, 让他们听到我们的声音!'

[Why is it that our students from the Jiangsu Providence must outscore Beijing students by many points to be admitted to the same college? This is a clear discrimination! The Commission of Education is the culprit! For the sake of our children's future, we must join force and mobilize, have dozens of our people sit-in at the Commission in Nanjing and make our voice heard!]

(2) State Critiques: Contents that are mainly arguing against a state policy that are mainly abstract academic discussions. Grievance must be generalizable to the entire country without an identifiable sub-group who is the most likely to voice such grievance. Examples of such category include national defense strategies, Constitutional amendments, and power dynamics among the political elites.

'主席任期终身制是严重的制度倒退, 与邓小平的改革思想背道而驰这种近乎独裁的做法, 不但不会带来党内民主, 反倒会让中国重新回到阶级斗争的泥沼之中, 发展停滞不前'

[The Constitutional Amendment that granted life tenure to the President is a serious institutional setback that directly contradicts with Deng Xiaoping's reforming ideas. Such measure will not only fail to bring internal democracy among the political elites but also put China back to the quagmire of class struggle that will stagnate our nation's development.]

(3) Historical Dispute (the control group): Contents that concerns historically sensible topics that have already been censored on the WeChat platform. The content is salvaged from website sources that are published outside of the Chinese internet domain. Examples of this category include the Tiananmen Square Incident, the Dalai Lama and Tibetan independence, and human rights practices in China.

'六四那天晚上天安门广场上死了至少数百学生学生和平抗议受到武装实弹镇压, 大量学生事后被问话, 处理'
[The night of the Tiananmen Square Incident costed lives of at least hundreds of college students. The peaceful protest was met with violent crackdown and live ammunition. Those who survived the night were later called in for questions and taken care of.]

(4) Scrambled text (the treated group): Contents that were originally texts from the third category but are then treated by having its syntax scrambled. The number of keywords and phrases remains the same, but the resulted text will be gibberish and cannot be immediately understood by a human reader.

'处理天安门广场受到大量被那天死了事后, 镇压六四数百上, 和平学生数百实弹武装抗议至少'
[Handling/ the Tiananmen Square/heavily killed/ by the day after/ survived/on/ the night/ hundreds of/ peace/ students/ live ammunition/ questioning/ armed protests/ at least/ taken care of/ violent crackdown]