Missing Links

A Comparison of Search Censorship in China

By Jeffrey Knockel, Ken Kato, and Emile Dirks

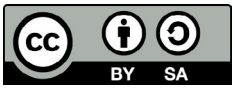
APRIL 26, 2023
RESEARCH REPORT #166

munk school
OF GLOBAL AFFAIRS & PUBLIC POLICY

UNIVERSITY OF TORONTO

THECITIZENLAB

Copyright



Electronic version first published by the Citizen Lab in 2023. This work can be accessed through https://citizenlab.ca/2023/04/a-comparison-of-search-censorship-in-china/.

Document Version: 1.0

About the Citizen Lab, Munk School of Global Affairs & Public Policy, University of Toronto

The Citizen Lab is an interdisciplinary laboratory based at the Munk School of Global Affairs & Public Policy, University of Toronto, focusing on research, development, and high-level strategic policy and legal engagement at the intersection of information and communication technologies, human rights, and global security.

We use a "mixed methods" approach to research that combines methods from political science, law, computer science, and area studies. Our research includes investigating digital espionage against civil society, documenting Internet filtering and other technologies and practices that impact freedom of expression online, analyzing privacy, security, and information controls of popular applications, and examining transparency and accountability mechanisms relevant to the relationship between corporations and state agencies regarding personal data and other surveillance activities.

Suggested Citation

Jeffrey Knockel, Ken Kato, and Emile Dirks. "Missing Links: A comparison of search censorship in China." Citizen Lab Report No. 166, University of Toronto, April 26, 2023. https://citizenlab.ca/2023/04/a-comparison-of-search-censorship-in-china/.

Acknowledgements

We would like to thank an anonymous researcher for contributing to this report. We would also like to thank Jedidiah Crandall, Jakub Dalek, Katja Drinhausen, Pellaeon Lin, Adam Senft, and Mari Zhou for valuable editing and peer review. Research for this project was supervised by Ron Deibert.

Contents

Key findings

- **Across eight China-accessible search platforms analyzed — Baidu, Baidu Zhidao, Bilibili, Microsoft Bing, Douyin, Jingdong, Sogou, and Weibo — we discovered over 60,000 unique censorship rules used to partially or totally censor search results returned on these platforms.**
- **We investigated different levels of censorship affecting each platform, which might either totally block all results or selectively allow some through, and we applied novel methods to unambiguously and exactly determine the rules triggering each of these types of censorship across all platforms.**
- **Among web search engines Microsoft Bing and Baidu, Bing's chief competitor in China, we found that, although Baidu has more censorship rules than Bing, Bing's political censorship rules were broader and affected more search results than Baidu. Bing on average also restricted displaying search results from a greater number of website domains.**
- **These findings call into question the ability of non-Chinese technology companies to better resist censorship demands than their Chinese counterparts and serve as a dismal forecast concerning the ability of other non-Chinese technology companies to introduce search products or other services in China without integrating at least as many restrictions on political and religious expression as their Chinese competitors.**

Introduction

Search engines are the information gatekeepers of the Internet. As such, search platform operators have a responsibility to ensure that their services provide impartial results. However, in this report, we show how search platforms operating in China infringe on their users' rights to freely access political and religious content, by implementing rules to either block all results for a search query or by only selectively showing results from certain sources, depending on the presence of triggering content in the query.

In this work we analyze a total of eight different search platforms. Three of the search platforms are web search engines, including those operated by Chinese companies — Baidu and Sogou — and one operated by a North American company — Microsoft Bing — whose level of censorship we found to in many ways exceed those of Chinese companies. While China's national firewall blocks access to web sites, the role that Baidu, Microsoft, and Sogou play in controlling information is in overcoming two of the firewall's limitations. First, due to the increasingly ubiquitous use of HTTPS encryption, China's firewall can typically only choose to censor or not censor entire sites as a whole. However, these search engine operators overcome this limitation by selectively censoring sites

depending on the type of information that the user is querying. Second, China's firewall operates opaquely, displaying a connection error of some kind in a user's web browser. By hiding the very existence of sites containing certain political and religious content, Baidu, Microsoft, and Sogou aid in preventing the user from being informed that they are being subjected to censorship in the first place.

We also examine search censorship on Chinese social media companies, namely Baidu Zhidao, Bilibili, Douyin, and Weibo. Perhaps more familiar to non-Chinese audiences are Douyin and Weibo. Douyin, developed and operated by TikTok's ByteDance, is the version of TikTok operating in China, and Weibo is a microblogging platform similar to Twitter. Perhaps less known are Baidu Zhidao and Bilibili. Baidu Zhidao is a question and answer platform similar to Quora operated by the same company as the Baidu search engine, and Bilibili is a video sharing site similar to YouTube. We also look at e-commerce platform Jingdong, which is similar to Amazon.

Given the strict regulatory environment which they face, users in China have limited choice in how they search for information. However, even among those limited choices, we nevertheless found important differences in the levels of censorship and in the availability of information among these search platforms. Most strikingly, we found that, although Baidu — Microsoft's chief search engine competitor in China — has more censorship rules than Bing, Bing's political censorship rules were broader and affected more search results than Baidu. This finding runs counter to the intuition that North American companies infringe less on their Chinese users' human rights than their Chinese company counterparts.

The remainder of this report is structured as follows. In "Background" and "Related work", we summarize the legal and regulatory environment in which Internet companies in China operate as well as existing research on Chinese search censorship. In "Model", "Methodology", and "Experimental setup", we describe how we model censorship rules, the manner in which we discover each platform's censorship rules, and the conditions in which we executed our experiments. In "Results", we reveal our findings of over 60,000 unique censorship rules being discovered, and we attempt to characterize which platforms censor more of what kinds of material. Finally, in "Limitations" and "Discussion", we discuss the limitations of our study, what our findings say about non-Chinese companies entering the Chinese market, and implications for future research.

Background

Internet companies operating in China are required to comply with both government laws concerning content regulations as well as broader political guidelines not codified in the law. Multiple actors within the government – including the Cyberspace Administration

of China and the Ministry of Public Security – hold companies responsible for content on their platforms, either through monitoring platforms for violations or investigating online criminal activity. Companies are expected to dedicate resources to ensure that all content is within legal and political or ideological compliance, and they can be fined or have their business licenses revoked if they are believed to be inadequately controlling content. China's information control system is characteristically one of intermediary liability or "self-discipline", which allows the government to push responsibility for information control to the private sector.

To understand the kind of information which is expected to be censored by companies in China, we can lean on at least four kinds of sources: (1) state legislation and regulations, (2) official announcements about state-led internet clean-up campaigns, (3) government-run online platforms where users can report prohibited material, and (4) official announcements about what kinds of prohibited material has been reported to the authorities.

Chinese government legislation and regulations have included provisions specifying what kinds of online content are prohibited. These documents include the Measures for the Administration of Security Protection of Computer Information Networks with International Interconnections (1997), the Cybersecurity Law (2017), Norms for the Administration of Online Short Video Platforms and Detailed Implementation Rules for Online Short Video Content Review Standards (2019), and Provisions on the Governance of the Online Information Content Ecosystem (2020). Many of the categories of prohibited content are shared among these four documents, as indicated in Figure 1. Shared categories include pornography and attacks on China's political system. However, it is also clear that more recent documents – in particular, the 2019 *Norms for the Administration of Online Short Video Platforms* and the 2020 *Provisions on Ecological Governance of Network Information Content* – have provided new categories of prohibited content. These include specific prohibitions against "harming the image of revolutionary leaders or heroes and martyrs" [损害革命领袖、英雄烈士形象] and more vague prohibitions against material which promotes "indecency, vulgarity, and kitsch" [低俗、庸俗、媚俗].

Measures for the Administration of Security Protection of Computer Information Networks with International Interconnections (1997)

Article 5:

No unit or individual may use the Internet to create, replicate, retrieve, or transmit the following kinds of information:

1. Inciting to resist or breaking the Constitution or laws or the implementation of administrative regulations
2. Inciting to overthrow the government or the socialist system
3. Inciting division of the country, harming national unification
4. Inciting hatred or discrimination among nationalities or harming the unity of the nationalities

5. Making falsehoods or distorting the truth, spreading rumors, destroying the order of society
6. Promoting feudal superstitions, sexually suggestive material, gambling, violence, murder
7. Terrorism or inciting others to criminal activity; openly insulting other people or distorting the truth to slander people
8. Injuring the reputation of state organs
9. Other activities against the Constitution, laws or administrative regulations

Cybersecurity Law (2017)

Article 12:

Any person and organization using networks shall abide by the Constitution and laws, observe public order, and respect social morality; they must not endanger cybersecurity, and must not use the Internet to engage in activities endangering national security, national honor, and national interests; they must not incite subversion of national sovereignty, overturn the socialist system, incite separatism, break national unity, advocate terrorism or extremism, advocate ethnic hatred and ethnic discrimination, disseminate violent, obscene, or sexual information, create or disseminate false information to disrupt the economic or social order, or information that infringes on the reputation, privacy, intellectual property or other lawful rights and interests of others, and other such acts.

Norms for the Administration of Online Short Video Platforms and Detailed Implementation Rules for Online Short Video Content Review Standards (2019)

4. Technical Management Regulations:

Based on the basic standards for review of online short video content, short video programs broadcast online, as well as their titles, names, comments, Danu, emojis, and language, performance, subtitles, and backgrounds, must not have the following specific content appear (commonly seen problems):

1. Content attacking the national political system or legal system
2. Content dividing the nation
3. Content harming the nation's image
4. Content harming the image of revolutionary leaders or heroes and martyrs
5. Content disclosing state secrets
6. Content undermining social stability
7. Content harmful to ethnic and territorial unity
8. Content counter to state religious policies
9. Content spreading terrorism
10. Content distorting or belittling exceptional traditional ethnic culture
11. Content maliciously damaging or harming the image of the state's civil servants such as from people's military, state security, police, administration, or justice, or the image of Communist Party members
12. Content glamorizing negativity or negative characters
13. Content promoting feudal superstitions contrary to the scientific spirit
14. Content promoting a negative and decadent outlook on life or world view and values
15. Content depicting violence and gore, or showing of repulsive conduct and horror scenes
16. Content showing pornography and obscenity, depicting crass and vulgar tastes, or promoting unhealthy and non-mainstream attitudes towards love and marriage

17. Content insulting, defaming, belittling, or caricaturing others
18. Content in defiance of social mores
19. Contents that is not conducive to the healthy growth of minors
20. Content promoting or glamourising historical wars of aggression or colonial history
21. Other content that violates relevant national provisions or social mores and norms

Provisions on the Governance of the Online Information Content Ecosystem (2020)

Article 6:

A network information content producer shall not make, copy or publish any illegal information containing the following:

1. Violating the fundamental principles set forth in the Constitution
2. Jeopardizing national security, divulging state secrets, subverting the state power, or undermining the national unity
3. Damaging the reputation or interests of the state
4. Distorting, defaming, desecrating, or denying the deeds and spirit of heroes and martyrs, and insulting, defaming, or otherwise infringing upon the name, portrait, reputation, or honor of a hero or a martyr
5. Advocating terrorism or extremism, or instigating any terrorist or extremist activity
6. Inciting ethnic hatred or discrimination to undermine ethnic solidarity
7. Detrimental to state religious policies, propagating heretical or superstitious ideas
8. Spreading rumors to disturb economic and social order
9. Disseminating obscenity, pornography, force, brutality and terror or crime-abetting
10. Humiliating or defaming others or infringing upon their reputation, privacy and other legitimate rights and interests
11. Other contents prohibited by laws and administrative regulations

Article 7:

A network information content producer shall take measures to prevent and resist the production, reproduction and publication of undesirable information containing the following:

1. Using exaggerated titles that are seriously inconsistent with the contents
2. Hyping gossips, scandals, bad deeds, and so forth
3. Making improper comments on natural disasters, major accidents or other disasters
4. Containing sexual innuendo, sexual provocations, and other information that easily leads to sexual fantasy
5. Showing bloodiness, horror, cruelty, and other scenes that causes physical and mental discomfort
6. Inciting discrimination among communities or regions
7. Promoting indecency, vulgarity, and kitsch
8. Contents that may induce minors to imitate unsafe behaviors, violate social morality, or induce minors to indulge in unhealthy habits
9. Other contents that adversely affect network ecology

Figure 1: Types of prohibited online content listed in government legislation and regulations.

The government legislation and regulations listed in Figure 1 are not the only official sources detailing what kinds of online content are either legally prohibited or are politically undesirable. Another indicator of what online material is censored are official descriptions of internet clean-up campaigns. Since 2013, China's cyber regulator the Cyberspace Administration of China, the Propaganda Department's Office of the National Working Small Group for "Combating Pornography and Illegal Publications", the Ministry of Public Security, and other party-state organs have conducted annual special operations for internet purification [净化网络环境专项行动, abbreviated as 净网]. These special operations involve identifying websites, platforms, and accounts which contain prohibited content, compelling the removal of content, and punishing those responsible through warnings or administrative or criminal penalties.

Internet purification operations initially concentrated on "obscene pornographic information" [淫秽色情信息]. But between 2013 and 2022, the focus of these special operations as stated in annual and semi-annual announcements widened to include a broader range of legally or politically prohibited content. An aggregate list of the targets of internet purification campaigns mentioned in these announcements is provided in Figure 2. Prohibited content mentioned in these announcements has recently included material which is "emotionally manipulative" [情感操控] (mentioned in 2020), "historical nihilistic" [历史虚无主义] (mentioned in 2021), or promotes "divination and superstition" [占卜迷信] (mentioned in 2022). An indication of the increasing breadth of these operations can be found in annual or semi-annual announcements made online by the Cyberspace Administration of China and the Ministry of Public Security about the progress of these operations. These announcements include information of the kinds of prohibited content which authorities have identified and removed. It is not clear if these annual announcements provide a full list of the material authorities targeted for removal during the year in question. Nonetheless, these announcements make clear that state-led internet purification campaigns routinely identify and remove not only pornographic online material, but many other kinds of prohibited content listed in relevant legislation.

- Political rumors; historical nihilism; misusing the 100th anniversary of the founding of the Communist Party to engage in commercial activity; tampering with the history of the Party and the nation; slandering heroes and martyrs; opposing basic Constitutional principles; information which threatens national security
- Violence; weapons; terrorism
- Harmful material related to ethnic groups and religion; promotion of heterodox faiths, feudal superstitions, and online divination
- Pornographic and vulgar content; socially harmful material; flaunting wealth and money worship; emotionally manipulative websites and platforms
- Gambling
- Fraud; illegal collection, editing, and publishing of financial information; publication of false information; blackmail; illegally buying and selling bank cards; sale of rare and endangered animals and plants

- Illicit drugs
- False advertising; false job recruitment posts; underhanded "black public relations"; paid internet posters
- False pharmaceutical information; sale of counterfeit drugs; copyright infringement and counterfeiting
- Illegal surrogacy; dishonest marriage websites
- Unauthorized providing of online news services; fake news; misleading or false information on the epidemic situation in Beijing
- Managing disorderly fan communities and online user accounts

Figure 2: Aggregate list of targets of special operations for internet purification mentioned in annual or semi-annual announcements, compiled from 2013, 2014, 2015, 2016, 2017, 2018, 2019, 2020, 2021, and 2022.

Beyond the targets of internet purification campaigns noted in Figure 2, further insight into what kind of online content is censored can be found on the Cyberspace Administration of China's Illegal and Undesirable Information Reporting Center [中央网信办⊠国家互联网信息办公室⊠违法和不良信息举报中心]. As part of the special operations for internet purification, the Cyberspace Administration of China encourages domestic internet users to make named or anonymous reports of prohibited "undesirable content" [不良信息内容] or "harmful information" [有害信息] through the Reporting Center. As part of the reporting process, users are asked to identify the kind of prohibited they have found according to nine categories provided by the Reporting Center, which are listed in Figure 3: politics, violent terrorism, fraud and blackmail, pornography, vulgarity, gambling, rights infringement, rumors, and a broadly defined category of "other." The nine categories listed in Figure 3 broadly match both the kinds of content proscribed under Chinese government legislation and regulations covering online content, as well as the targets of internet purification special operations listed in announcements made by the Cyberspace Administration and the Ministry of Public Security.

- Politics
- Violent terrorism
- Fraud and blackmail
- Pornography
- Vulgarity
- Gambling
- Rights infringement
- Rumors
- Other, including: online borrowing and lending; online criminal activity; online commercial disputes; online false and illegal advertising; online extortion and post deletion; email and telephone harassment; intellectual copyright infringement; piracy; fake media and fake journalists; gang activity; online cultural market activity including music, performance, and animation; and telecommunications user services

Figure 3: Categories of prohibited material listed by the Cyberspace Administration of China's Illegal and Undesirable Information Reporting Center.

Online announcements made by the Cyberspace Administration of China about the number of reports of prohibited content also suggest the kinds of content the Chinese state seeks to censor. These announcements arrange prohibited content into various categories, which only partially match those listed by the Reporting Center. Over the years these subcategories have included pornography, politics, vulgarity, gambling, rights infringement, rumors, terrorism, fraud, online extortion, paid post deletion, and other forms of content.

We performed searches on Google and Baidu for terms associated with these announcements — "全国网络举报受理情况" [national situation of the handling of online reports] and "全国网络举报类型分布" [national distribution of categories of online reports] — during February and March 2023. We collected websites which provided breakdowns of the categories reports of prohibited content made by Cyberspace Administration offices across China ("全国各地网信办") and specific websites ("各网站"). We found some of these announcements on websites run by the national Cyberspace Administration of China or reporting websites run by provincial authorities, while others were published on news media websites.

These announcements do not appear to be consistently available, and we were only able to find ten announcements: nine monthly announcements released between January 2016 and January 2017, and one annual announcement for the year 2020, which are presented in Table 1. Nonetheless, the announcements which are available indicate the kinds of material that are reported and censored across platforms and websites in China. In addition, these announcements provide statistical information on the number of reports of prohibited content per category. We have provided a breakdown of this statistical data in Table 1. Based on the statistical data contained in these announcements, the majority of reported prohibited content is pornographic, followed by either political content or "other" material which does not fall into any of the other categories.

	2016 Jan	2016 May	2016 June	2016 Aug	2016 Sept	2016 Oct	2016 Nov	2016 Dec	2017 Jan	2020
Pornography	64.7	60.4	60.4	60.7	63.7	60.3	60.7	50.0	55.2	61.7
Politics	8.9	12.9	11.3	11.9	11.8	13.1	13.9	29.0	23.5	7.7
Vulgarity	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	3.3
Gambling	1.4	1.3	1.8	3.9	2.2	3.5	1.8	1.6	1.6	9.8
Rights Infringement	2.5	5.7	3.9	2.7	3.9	4.0	4.3	3.9	4.1	2.2
Rumors	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	1.1
Terrorism	0.1	2.1	0.7	2.0	1.0	2.1	1.1	1.1	1.1	0.9
Fraud	4.5	8.2	11.9	8.4	7.1	7.1	8.0	4.9	5.0	1.3
Online Extortion and Paid Post Deletion	0.2	0.1	0.2	0.6	0.8	0.1	0.5	0.5	0.5	n/a

	2016 Jan	2016 May	2016 June	2016 Aug	2016 Sept	2016 Oct	2016 Nov	2016 Dec	2017 Jan	2020
Other	17.7	9.3	9.8	9.8	9.5	9.8	9.7	9.0	9.0	12.0

Table 1: For announcements spanning 2016 to 2020, the % of reports in that announcement spanning each prohibited online content category. Archived copies for these announcements are linked through the respective date.

Other Chinese government announcements give an indication of which platforms are responsible for hosting prohibited content. These reports provide monthly or annual totals of the number of pieces of prohibited content reported to the authorities, broken down according to the website or platform on which the content was found. Alongside warning, fining, or in other ways punishing companies for hosting prohibited content, these public reports have the function of naming and shaming companies for failing to fully comply with Chinese laws on online content management.

Platform	# of Reports
Weibo	53.126 million
Baidu	25.961 million
Alibaba	11.689 million
Kuaishou	6.59 million
Tengxun	6.309 million
Douban	3.514 million
Zhihu	2.143 million
Jinri Toutiao	2.063 million
Sina Wang	934,000
Sogou	331,000

Table 2: For different Internet platforms, the number of reports of prohibited content for 2021.

The most recent announcement we found concerning reports of prohibited content broken down by platform is for the year 2021. The statistical data contained in this announcement, presented in Table 2, indicates that Weibo was subject to the majority of reports of prohibited content with 53.126 million reports, followed by Baidu (25.961 million) and Alibaba (11.689). The ten platforms listed in Table 2 account for roughly 110 million reports. According to the government announcement from which these data come, these 110 million reports are 75.6% of the 166 million reports of online prohibited content made in 2021.

Related work

There is a large body of previous research analyzing search platform censorship in China. Much of the earliest work focused on comparing censorship across web search engines accessible in China. In 2006, Reporters Without Borders tested by hand six keywords

across multiple search engines accessible in China, finding that Yahoo returned the most pro-Beijing results among the first ten results compared to other search engines. In the same year, Human Rights Watch tested by hand 25 keywords and 25 URLs, finding that Baidu and Yahoo were the most censored. This earliest work was limited in analyzing keyword-based censorship by attempting to characterize and compare the top *n* results for a searched keyword. This type of analysis is limited due to its subjectivity and its inherent assumption that the search engine with the most politically sensitive results must be the least censored when there may be other explanations for fewer sensitive results than the application of censorship rules.

In 2008, in a follow up to the previous studies, Citizen Lab researchers tested 60 hand-picked keywords across multiple search engines using an approach in which search queries were formed by combining a keyword with a web domain preceded by the "site:" operator to determine which domains were censored from the results of which keywords. Of all of the previous work we review, this work is the closest to ours. However, there are nevertheless fundamental differences. The work, like its predecessors, was limited to small sample sizes and relies on hand-picked samples. More importantly, its methods also cannot differentiate between a search query which is censored and one that genuinely has no results. While this may not seem significant in the context of testing hand-picked keywords, in our work we develop a method which can test whether a string of text triggers a censorship rule even when it would ordinarily return no results, and our method can isolate the exact keyword or keywords present in that string which are triggering its censorship. This capability was necessary for bridging the gap between testing lists of curated keywords versus testing long strings of arbitrary text, which is a necessary component for automated and ongoing testing of strings of text from sources such as news articles.

In a 2011 work, to instrument automated and ongoing censorship testing, Espinoza et al. developed a novel method using named entity extraction to select interesting keywords from a long string of news article text to use in search engine testing. Specifically, their method was designed to extract certain nouns, namely, the names of people, places, and organizations. The significance of this work is that it facilitates automatic censorship testing which does not rely on hand curation of keywords but instead can take as input long strings of arbitrary text, such as from news articles, automatically selecting from that text certain keywords to test. However, it is limited in that it makes assumptions about what type of content is likely to be sensitive, namely, certain kinds of nouns, and was used to test keywords for censorship individually. In contrast, we have found that censorship rules often require the presence of multiple, typically related keywords and commonly consist of a variety of parts of speech. Instead of selecting interesting keywords from a long string of text to test individually, our method tests a long string of text for the presence of censored content as a whole, even if it would not otherwise

have any search results. We can then, using additional search queries, isolate the exact keyword or keywords triggering the censorship of that text. Our method is completely agnostic to and requires no assumptions concerning parts of speech, semantics, word boundaries, or other aspects of language and effortlessly generalizes to Chinese, English, Uyghur, and Tibetan languages, among others.

In another 2011 work, Zhu et al. use automated methods to test curated keywords consisting of the 44,102 most-searched keywords on Baidu and Google.cn, 133 keywords known to be censored by China's national firewall, 1,126 political leaders of the Chinese government, and 85 keywords chosen by hand based on current events. This work is to our knowledge the first to speculate about the existence of different "white lists" of domains allowed to appear in the results for censored queries, whereas previous work has been framed in measuring which domains were blocked. In our work, we confirm the existence of these lists of authorized domains and attempt to quantify how many different lists exist, characterize when each list is applied, and measure which domains appear on each one.

More recently, in 2022 Citizen Lab researchers analyzed Microsoft Bing's autosuggestion system for Chinese-motivated political censorship, finding that not only was it applied to users in mainland China but that it was also applied, partially, to users in North America and elsewhere. While this work is unlike ours in that it analyzed for Chinese censorship queries' autosuggestions as opposed to the queries' results proper, it is related to our work in that it studies the censorship of Bing, the only remaining major non-Chinese web search engine accessible in China.

Most recent work studying search platform censorship in China has analyzed the search censorship performed by social media platforms, namely that of Chinese microblogging platform Sina Weibo. For instance, in 2014, as part of the ongoing Blocked on Weibo project, Ng used automated methods to test for the censorship of 2,429 politically sensitive keywords previously curated by China Digital Times, finding that 693 were censored with explicit notifications. In a follow-up study months later, Ng found that most of these no longer had explicit censorship notifications but still returned zero results. Ng speculated that this may be due to either the removal of keywords from search censorship which were still being applied to post deletion censorship or due to Weibo transitioning to a more covert form of censorship. Our findings in this report suggest that both hypotheses for his findings could be true, as in Appendix A we demonstrate a method for evading Weibo search censorship but which often still yields zero results due to a simultaneous application of search censorship rules to post deletion, but also much of our analysis of Weibo focuses on a form of soft censorship which subtly restricts which results can appear for sensitive queries.

Often work looking at Weibo censorship is ad hoc and non-methodological, performed quickly in response to ongoing events, often to be featured in news articles or social media. For example, in 2017, Citizen Lab researchers studied Weibo search censorship of human rights advocate Liu Xiaobo leading up to and in the wake of his passing. They found that censorship of his name and surrounding topics intensified immediately following his passing but eventually returned to baseline levels. In our work we facilitate a method to automatically and methodologically detect search censorship rules introduced in response to developing news events with the intention to aid such rapid investigations.

In response to a 2022 incident in which Canadian Prime Minister Justin Trudeau had a heated, public conversation with Chinese President Xi Jinping, journalist Wenhao Ma tweeted his discovery that "特鲁多" [Trudeau], "小土豆" [little potato] (a Chinese nickname for Trudeau), and the English word "potato" were censored by Weibo search. We highlight this example for two reasons. First, Ma identified these keywords as being censored even though they had search results because their search results seemed to only contain results from official accounts with blue "V" insignia, recognizing that Weibo was applying a more subtle, softer form of censorship compared to simply displaying zero results. In our work, we develop a method to measure unambiguously when such keywords are subject to this type of censorship without attempting to glean it from the number of results from official accounts. Second, however, Ma's claim that the English word "potato" was censored by Weibo was, while correct, misleading in that its censorship had nothing to do with Trudeau or potatoes but because it contains the substring "pot", a slang term for marijuana. To avoid this type of inadvertent misattribution, in our work, we use a carefully designed algorithm to extract the exact keywords or combination of keywords triggering the censorship of a string of text.

Model

In previous work studying automatic censorship of messages on WeChat, we determined that WeChat automatically censors messages if they contain any of a number of blocked keyword combinations, and we had defined a keyword combination as a set of one or more keywords such that, if each keyword in the combination was present somewhere in a text, the text would be censored. For instance, if WeChat censors the keyword combination {"Xi", "Jinping", "gif"}, then any message containing all of these keywords, anywhere in the message, in any order, is censored. Thus, this combination would censor "Xi Jinping gif" and "gif of Xi Jinping" but not "Xi gif". In this model, keywords can overlap as well, so even "Xi Jinpingif" would be censored since it contains the strings "Xi", "Jinping", and "gif" somewhere in the message, although the latter two overlap.

To our knowledge, this manner of modeling WeChat's automated chat censorship rules as a "list of unordered sets" of blocked keywords completely captured WeChat's censorship behavior. However, other censorship systems fitted with different censorship implementations may not be able to be adequately modeled using this model. For instance, a "list of ordered sequences" censorship system might require that the keywords appear in a specific order. For example, the rule ("Xi", "Jinping", "gif") would censor "Xi Jinping gif" but not "gif of Xi Jinping". A censorship system which implemented rules as a series of regular expressions would not only require the keywords to appear in an order but also that they not overlap. For example, the regular expression /Xi.*Jinping.*gif/ would censor "Xi Jinping gif" but neither "gif of Xi Jinping" nor "Xi Jinpingif". Finally, a censorship system might use some machine learning algorithm to classify which queries to censor. However, we have not previously observed such systems used to perform real-time, political censorship, likely due to the requirements of such a system to operate with low false positives and to possess a nuanced, day-by-day understanding of what content is politically sensitive.

To attempt to capture the censorship behavior of as many search platforms as possible, in the remainder of this work we chose to use a "list of ordered sequences" model as in doing so we are being as conservative in our assumptions as possible. For instance, by using ordered sequences, we can still model unordered rules, although this may require multiple ordered sequences to capture every possible permutation (e.g., ("Xi", "Jinping", "gif"), ("gif", "Xi", "Jinping"), etc.). In our model we allow for the possibility that keywords triggering censorship in a query may be overlapping, but by facilitating this possibility we can still measure systems where keywords cannot.

Throughout the remainder of this work, we use the term *keyword combination* to refer to such a "list of ordered sequences", and we will express them as keywords separated by plus signs, e.g., "Xi + Jinping + gif". Later in our work, we reflect on this model more. In our "Methodology" section, we explain exactly how we measure which sequence of keywords is triggering the censorship of a censored query, and in our "Results" section we reflect on how effective our model performed in capturing the actual censorship behavior of the search platforms which we measured.

Methodology

In this section we describe our overall experimental methodology and then detail the methodologies of three different experiments which we perform.

Search platforms analyzed

We aimed to analyze the most popular platforms across different kinds of Internet platforms ranging from web search engines and e-commerce platforms to social media sites. Overall, we selected to analyze eight different search platforms, including three web search engines, four varying types of social media network, and one e-commerce platform (see Table 3 for the full list).

Website	Description	Object of Censorship
Baidu	Web search engine	Web page results
Baidu Zhidao	Q&A platform	Q&A post results
Bilibili	Video sharing platform	Video results
Bing	Web search engine	Web page results
Douyin	Operated by TikTok's ByteDance, the version of TikTok accessible from mainland China	Video results
Jingdong	E-commerce platform	Product recommendations
Sogou	Web search engine	Web page results
Weibo	Microblogging site similar to Twitter	Microblog results

Table 3: The search platforms that we analyzed and the object of censorship which we measured on them.

Notably, our selection includes one platform not operated by a Chinese company — Microsoft Bing — whereas the remaining are all operated by Chinese companies.

Measuring whether a query is censored

When a user using a search platform searches for a query, this query is sent to a server. If the query is not censored, the server will respond with the corresponding matches to the query. However, with a censored query, there are two possibilities depending on the search platform:

1. The server returns a unique notification when the user's query contains sensitive content. We call this *transparent* censorship because the signal is unambiguous.
2. The server spuriously omits some or all search results despite that content matching the user's query. We call this *opaque* censorship due to there existing an ambiguity as to whether the query was censored or whether those matches never existed.

For search platforms which employ transparent censorship, measuring whether a query is censored is straightforward: test the query and check if there is a notification that the query is censored. However, for search platforms which censor opaquely, we were required to employ a more sophisticated methodology to distinguish between cases

where there are genuinely zero matches and cases of opaque censorship. In the following section we discuss the method we used to distinguish between these cases.

Measuring whether a string of text is opaquely censored

On platforms which employ opaque censorship, in order to distinguish between cases where there are genuinely zero matches and cases where matches exist but are being opaquely censored, we use a technique of creating test queries for a string of text such that they should always return matches unless the string of text is censored, tailored to each platform. We call such a modified query a *truism*, which Wikipedia defines as "a claim that is so obvious or self-evident as to be hardly worth mentioning, except as a reminder or as a rhetorical or literary device". Our truisms are search queries which should obviously return results but are used as devices to unambiguously detect the presence of censorship of a string of text.

As an example, on Baidu Zhidao, we create a truism by surrounding the string in question with "the -(" before the string and ")" after the string. Thus, for Baidu Zhidao, to test the string "习近平" we would test the truism "the -(习近平)". On Baidu Zhidao this syntax indicates to the search platform to logically negate whatever is in between the parentheses and can be interpreted as searching for all results containing "the" which do *not* contain "习近平". In the case of Baidu Zhidao and many other platforms, we have discovered that even content which is negated in a query can still trigger the query's censorship.

Website	Transformation	Explanation
Baidu	"site:com.cn␣-(" + *string* + ")"	Since there is no character or string which is present on every web page, we negate the possibly-censored string and since Baidu queries cannot begin with a negation we use the "site:" operator to restrict results to a popular top-level domain.
Baidu Zhidao	"the␣-(" + *string* + ")"	Since posts exist containing "the", this query requesting pages containing either "microsoft" or the possibly-censored string should always return results unless the string is censored.
Bilibili	*string*	Transparent censorship (although Bilibili does not report a censorship notification, a unique error code is returned by Bilibili's API although this is invisible during ordinary usage)
Bing	"microsoft␣|␣" + *string*	Since web pages exist containing "microsoft", this query requesting pages containing either "microsoft" or the possibly-censored string should always return results unless the string is censored.

Website	Transformation	Explanation
Jingdong	*string*	Transparent censorship (although Jingdong does not report a censorship notification, the platform only fails to provide any recommendations when a query is sensitive)
Sogou	"site:com.cn#龘" + *string*	Although Sogou does not support disjunction (X | Y) or negation (-X), we discovered that Sogou supports a "site:" operator restricting results to the proceeding URL's domain. While only the domain of the URL is used to restrict results, we found that other elements of a URL, such as a path or fragment may nevertheless be included although they have no effect on the results of the query. Therefore, by searching for "site:com.cn#XXX", we are merely searching for any page with a top-level "com.cn" domain, and the fragment proceeding the "#" only exists to trigger censorship. Finally, since Sogou's censorship system strips punctuation including hashes from queries before testing them for sensitive content, we place a "龘", an extremely rare Chinese character unlikely to appear in censored content, between the "#" and the possibly-censored string to cause the censorship system to not join the possibly-censored string to "com.cn". The "龘" character was also chosen because, unlike punctuation which causes Sogou's URL parser to stop parsing the text as a URL, a "龘" is allowed by Sogou's parser to be in a URL fragment.
Weibo	*string*	Transparent censorship (Weibo displays a censorship notification)

Table 4: Rules for transforming a query for string *to a truism testing for censorship of* string *such that* string *is censored if and only if its corresponding truism returns zero results.*

As another example, while Baidu Zhidao and many other platforms seemed to naively scan queries for the presence of strings to trigger censorship, Bing's censorship system seemed clever enough to not allow the content of negated content to trigger censorship. However, we were still able to create a truism on Bing by searching for "(microsoft | 习近平)". On Bing this syntax indicates to the search platform to return results that contain either "microsoft" or "习近平". Since we know that there exist pages on the Internet containing "microsoft" and since "microsoft" is not censored, then, if there are no results, it must be because "习近平" is censored. See Table 4 for the rules which we used to create truisms to test each site employing opaque censorship.

Although we could theoretically construct such queries, note that truisms are not necessarily *tautological*, i.e., they are not guaranteed to return results *a priori*. For instance,

we could construct a query "(习近平 | -习近平)" which would request any result that either contains 习近平 or does not contain 习近平 (i.e., every result). However, in our testing search platforms did not seem designed to recognize such queries as tautologies and often the results would be logically inconsistent (e.g., "习近平" reporting more results than "(习近平 | -习近平)"). As such, by "truism" we refer to queries which when not censored are merely certain to return results in practice although not necessarily *a priori*.

Since whitespace and punctuation characters can induce unpredictable behavior on censorship systems and because it can potentially interfere with the syntax added by our truisms, we strip all whitespace and punctuation from strings before testing. While it is possible that by performing this practice we may be failing to discover some censored rules which require punctuation, we found that in our previous study of WeChat that WeChat strips whitespace and punctuation from messages before testing them for censorship and that failing to strip these characters ourselves resulted in the spurious inclusion of them in our results. Therefore, out of caution, because we prefer accuracy of our results over the possibility of a slightly larger size of results, we strip whitespace and characters from our test strings before testing.

By using the method described in this section of testing using truisms, we can be certain that if our query does not have any returned matches then it must be due to the result of censorship. Thus far, we have discussed how we measure *hard* censorship, that is, censorship which denies the user from any matches. However, in this following section, we discuss how to measure a more subtle form of censorship in which the matches may be partially censored.

Measuring whether a query is partially censored

Across the three web search engines we tested, many queries which did return results only returned results linking to web sites which were Chinese state-owned or state-approved media outlets (see Figure 4 for an illustration).

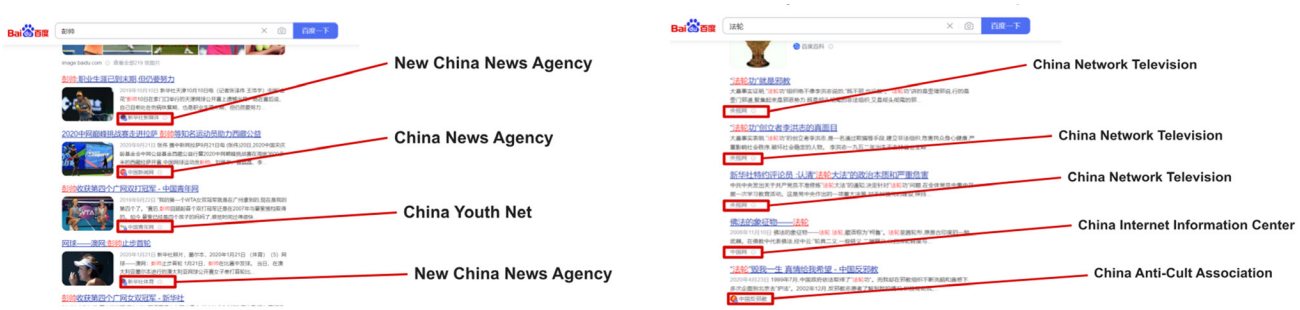


Figure 4: On Baidu, an example of a query whose results are only from Chinese state media.

Moreover, for some social media search platforms, we noticed that, for some queries that did return results, these results seemed to be only from accounts which have received a

certain amount of verification or approval. We call this type of censorship in which results are only allowed from authorized sources *soft* censorship and censorship in which no results are allowed *hard* censorship (see Table 5 for a breakdown of each platform we discovered performing soft censorship).

Website	Soft censorship
Baidu	Only shows results from authorized domains (typically Chinese government sites, Chinese state media, etc.)
Bing	Only shows results from authorized domains (typically Chinese government sites, Chinese state media, etc.)
Douyin	Only verified accounts
Sogou	Only shows results from authorized domains (Chinese government sites, Chinese state media, etc.)
Weibo	Only verified accounts

Table 5: Platforms which we discovered performing soft censorship and the manner in which they perform it.

To detect this form of soft censorship, for each web search engine, we modified its truism by restricting the results to only be allowed from unauthorized sources. For example, on Baidu, we only allow results from microsoft.com, a site we chose because it is both popular and accessible in China but foreign operated and unlikely to be pre-approved for voicing state propaganda. For Baidu, we surrounded the tested string with "site:microsoft.com -(" on the left and ")" on the right in order to transform it into a truism and test it for soft censorship but with the restriction that results were only allowed from an unauthorized source. Thus, for the string "彭帅", we would test the truism "site:microsoft.com -(彭帅)", which can be interpreted as searching for any page on microsoft.com not containing "彭帅". See Table 6 for the rules which we used to create truisms to test each site employing soft censorship.

Website	Transformation	Explanation
Baidu	"site:microsoft.com␣-(" + *string* + ")"	Same as in Table 4 except restricted to microsoft.com, an unauthorized site.
Bing	"site:microsoft.com␣(microsoft␣|␣" + *string* + ")"	Same as in Table 4 except restricted to microsoft.com, an unauthorized site.
Douyin	*string* + "␣‰‰"	Douyin normally displays results for any query, no matter if there exist no results which are an exact match or even a close match. However, when queries are soft censored, Douyin applies two restrictions to results, namely that results must contain all words in the search query and that results must be from verified accounts. As such, we additionally search for "‰‰", which has not been posted by any verified account, so that soft-censored queries will never display search results.
Sogou	"site:microsoft.com#" + *string*	Same as in Table 4 except restricted to microsoft.com, an unauthorized site.

Website	Transformation	Explanation
Weibo	“‰‰␣-(” + *string* + “)”	Only non-verified accounts have posted the string “‰‰” and there is no character or string which is in all posts containing “‰‰”.

Table 6: Rules for transforming a query for string *to a truism testing for soft censorship of* string *such that* string *is soft censored if and only if its truism returns zero results.*

Notably, although many sensitive queries on Douyin returned zero results, we did not find any evidence of hard censorship on Douyin that could not be explained by the soft censorship system which we explain in Table 6. As such, on Douyin, we only measure its soft censorship system.

Isolating which keywords are triggering censorship

Thus far we have discussed how to determine whether a string of text is either hard or soft censored across each of the search platforms which we tested. However, given that a string of text is censored, we still desire to know which keyword or combinations of keywords present in that text are responsible for triggering its censorship. In this section, we outline the method we employ of *isolating* which combination of keywords is triggering a text’s censorship by making additional queries.

To isolate keywords triggering censorship of a string of text, we make use of an algorithm called CABS which we originally introduced in 2019 and continue to maintain here. Our original algorithm was motivated by discovering censored keyword combinations on WeChat, which we modeled as a “list of unordered sets”, but, as we model censorship in this work as a “list of ordered sequences”, we adapted the algorithm to fit this model. Fortunately the changes required were trivial, essentially replacing all sets and set operations with tuples and their corresponding tuple operations (e.g., set union is replaced with tuple concatenation). To fully understand the algorithm and to access its code, we recommend visiting the previous links in this paragraph. However, in the remainder of this section we will briefly outline the intuition behind how this isolation algorithm works.

The algorithm works by performing bisection and attempting to truncate as much of the text being isolated at a time while preserving the property that it is still censored. For example, initially it will attempt to remove the second half of the text and measure if it is still censored. If it is, then it will attempt to remove the last three quarters of the text. If it is not, then it will attempt to remove only the last quarter of the text. By iteratively repeating this procedure, the algorithm eventually discovers the final character of one of the keywords triggering the censorship of the string. It next attempts to discover the first character of this keyword. Once the complete keyword is discovered, the algorithm tests if the keyword it has discovered thus far is sufficient to trigger censorship. If not, it repeats this process of finding another keyword in the censoring keyword combination on the remaining text up to but not including the final character of the keyword most

recently discovered. It repeats this process until enough keywords have been discovered to trigger censorship, producing the censored keyword combination in its entirety.

There are, as it turns out, many subtleties to doing this correctly and efficiently, especially when keywords can overlap or when there may be present multiple keyword combinations triggering censorship. However, through careful design and testing, our algorithm is correct even in the presence of such corner cases.

Website	Character	Explanation
Baidu	"‰"	Efficiently encoded in GB18030, the encoding under which our testing query length on Baidu is limited.
Baidu Zhidao	"‰"	Efficiently encoded in GBK, the encoding under which our testing query length on Baidu Zhidao is limited.
Bilibili	"😀"	Sufficient to separate keywords.
Bing	"😀"	In our testing, we found that different join characters could produce different results, suggesting that Bing may use complicated rules for how keywords in a single rule may be separated. In our report, we use "⊠" as our "join" character, although another choice may have split keyword combinations into a greater number of keywords.
Douyin	"😀"	Sufficient to separate keywords.
Jingdong	"␣"	Sufficient to separate keywords.
Sogou	"龖"	Efficiently encoded in GB18030, the encoding under which our testing query length on Sogou is limited, and, unlike punctuation or other characters such as "‰", can be parsed by Sogou as part of a URL fragment (see Table 4). We chose this specific Chinese character because it is extremely rare (refers to a stuffy, nasal voice) and thus unlikely to collide with actual censorship rules but occurs in the basic multilingual plane.
Weibo	"‰"	Efficiently encoded in GB18030, the encoding under which our testing query length on Weibo is limited.

Table 7: For each tested search platform the "join" character that is used when isolating the combination of keywords triggering the censorship of a string of text.

The one way in which the algorithm must be adapted to a given search platform is by choosing a "join" character. This selection is necessary as not every platform considers the same characters as splitting a keyword. For instance, on one platform, putting spaces in between the characters of a censored keyword may not prevent it from being censored but on another it may. A desirable "join" character for a platform is one whose insertion into a censored keyword would prevent it from censoring but also one that it is unlikely to appear in censored keywords that we wish to measure and that can be encoded efficiently in whatever character encoding a search platform internally uses. For a breakdown of the "join" characters that we used for each tested search platform with corresponding motivations, please see Table 7.

By using this algorithmic technique, we can determine the exact keyword or combination of keywords necessary to trigger censorship of a censored string of text. The results are not statistical inferences, approximations, or in any way probabilistic. Moreover, the algorithm is agnostic of and makes no assumptions concerning language, including concerning parts of speech, semantics, word boundaries, or other aspects of language, and it effortlessly generalizes to Chinese, English, Uyghur, and Tibetan languages, among others.

Overcoming testing hazards

During the preliminary design and testing of our methods, we observed that our methodology would need to overcome multiple hazards in order to provide thorough and accurate results, including captchas, various restrictions on query length, and inconsistent results returned by search platforms. Below we outline how we overcome these hazards.

Captchas

We found that, after a period of automated testing, Sogou and Baidu began displaying captchas instead of displaying search results until the captchas were solved. We did not investigate attempting to solve the captchas automatically. However, for Sogou, we found that whenever presented with a captcha we could restart the browser session to resume testing for some substantial period of time until the next captcha appeared, at which point we could simply restart the browser session again. For Baidu, restarting the browser session was typically ineffective. However, we found that solving a captcha would allow a browser session to test for 24 hours uninterrupted by captchas. Thus, to instrument Baidu's testing, every 24 hours we manually solve a captcha for each browser session, requiring only seconds of manual intervention each day. However, if Baidu's captcha displays became more frequent or if we wanted to completely automate the testing, future work might look at applying software designed to automatically solve these captchas.

Douyin also displayed captchas. However, unlike with Sogou and Baidu, even after solving Douyin's captchas, repeated search querying would inevitably begin yielding zero results for any search query, regardless of its sensitivity. As such, we were not able to complete every experiment with Douyin, as we stopped testing it early in our analysis due to this limitation.

Query length limitations

The search platforms we tested have limitations in the length of query which we could test. Exceeding these limits had various consequences, such as the platform returning an error message, silently truncating the query, or all content beyond the limit evading censorship. As such, for each platform, we performed testing to determine the value of any applicable limit. As characters can take varying space in different representations or encodings, we also had to determine the unit of the limit, which we found to vary across platforms as being a function of the number of raw Unicode characters or the number of

bytes in some character encoding such as UTF-8, UTF-16, GB18030, etc. (see Table 8 for a complete breakdown).

Website	Maximum query length	Encoding and unit
Baidu	76	GB18030 bytes
Baidu Zhidao	76	GBK bytes
Bilibili	33	Unicode characters
Bing	150	UTF-8 bytes
Douyin	202	UTF-16 bytes
Jingdong	80	Unicode characters
Sogou	80	GB18030 bytes
Weibo	40	GB18030 bytes

Table 8: For each search platform, the maximum query length we used in testing.

Our code was written to ensure that queries were never tested which exceeded a platform's limits to ensure the reliability of our results.

Inconsistent search results

We observed inconsistencies in the search results with some search engines during our testing. When we searched for a truism for the first time, we found that some platforms would occasionally return no results for a truism, even if it is not censored. Testing it again would yield results. We hypothesize that the eccentric queries which we construct would sometimes overwhelm the search platform but, once it had sufficient time to be primed, it would then return results for subsequent searches using that query. For other platforms, we also observed cases in which it seemed that we would see a small number of censorship rules being applied inconsistently. We hypothesize that such inconsistent observations may have resulted from load balancing between servers with small differences between the censorship blocklists with which they had been deployed. In any case, to make our measurements robust to these and other inconsistencies, we apply the following algorithm, expressed below in Python-like pseudocode, which effectively retests a potential keyword combination an additional two times over the span of three hours before considering it a censored keyword combination:

```
1: def robust_isolate(censored_text):
2:     combo = isolate(censored_text) # returns list of keywords
3:     for round in range(2): # retest an additional two times
4:         wait_for_an_hour()
5:         last_combo = combo
6:         censored_text = ‘’.join(combo)
7:         combo = isolate(censored_text) # returns list of keywords
8:         if combo != last_combo:
```

```
9:             if len(combo) < len(last_combo) or (len(combo) == len(last_
combo) and len(‘’.join(combo)) < len(‘’.join(last_combo))): # force a measure
of progress to ensure termination of algorithm
10:                 return robust_isolation(‘’.join(combo)) # restart with new
11:             return None # give up
12:     return combo
```

In this code, in the event that we discover an inconsistent result, we do one of two things depending on how the new result compares to the previous one. If, compared to the previous result, the new result's keyword combination has either fewer constituent keywords or if it both has the same number of constituent keywords but the sum of the lengths of each of its constituent keywords is less than those of the previous result, then we restart the robust isolation process from scratch on the new keyword. Otherwise, we simply give up attempting to isolate the triggering keyword combination from the given censored string. We have this rule in place to ensure that the isolation of the keyword combination is making some measure of progress, in either having fewer keywords or in having the same number of keywords but shorter ones. This policy ensures that, in an environment where servers may be giving inconsistent results, the algorithm still terminates, either by eventually returning a reliable result or by failing. Although we did not collect data specifically pertaining to this matter, we believe from casual observation that such failures are exceedingly rare and occur only when nothing else could have been easily done to obtain a reliable result.

Experiments

In this work we perform three experiments using different sampling methodologies to address different research questions that we had. In our first two experiments, we test search platforms for the censorship of people's names and of known sensitive content, respectively. We also present a third, ongoing experiment from which we already have preliminary results, in which we test text for censorship from daily news sources. In the remainder of this section we set out the design of these three experiments in greater detail.

Experiment 1: Measuring censorship of people's names

In our first experiment we test people's names. Individuals or their names have the following desirable properties:

- Individuals can represent highly sensitive or controversial issues.
- Unlike more abstract concepts, a comprehensive sample of notable people and their names can be automatically curated and enumerated into a large test list.
- As opposed to a list of handpicked keywords or a list of sensitive keywords censored in other Chinese products, a list of people's names is not biased toward the sort or style of keywords censored in other Chinese products or toward a researcher's preconceptions.

To facilitate this experiment, we used a list of 18,863 notable people whose names which we had previously curated from Wikipedia in 2022. The manner in which we curated these is spelled out in a previous report, but, at a high level, these names were collected from Wikipedia by looking for people whose articles had a sufficiently high number of Wikipedia views and whose names had a sufficiently high amount of search volume on Microsoft Bing. While this list of notable names inevitably contained the names of famous Chinese politicians, political dissidents, and others whom we might expect to be the targets of censorship, the criteria through which we selected these names was designed to be unbiased and to also produce names for testing whose censorship we might not expect with the intention that we find surprising results.

In this experiment we test each person's name individually. For each name on this list, to generate a test string, we take the person's name as expressed in the Wikipedia article title and append, if different, the name in simplified Chinese characters and append, if different, the name in traditional Chinese characters, forming a final test string of between one and three variations of the name concatenated together. If in testing we find that the test string is censored, we then use our isolation algorithm to isolate a keyword combination triggering its censorship.

While isolating the triggering keyword combination may not seem necessary when individually testing keywords such as people's names, as it might seem apparent that sensitivity of that person's name must be responsible for triggering the censorship, we found it helpful in discovering cases where names were collaterally censored, either by accidentally containing another part of another censored name (e.g., "习" [Xi]), or by accidentally containing other sensitive characters triggering its censorship (e.g., "伍富橋" [Alvin Ng] censored due to containing the character "橋" [bridge] following the Sitong Bridge Protests).

During this experiment, for each platform tested, we record each censored name and the keyword combination triggering its censorship.

Experiment 2: Measuring censorship of known sensitive content

In our second experiment we test from a compilation of known sensitive content. Previous work has shown that, to comply with Chinese censorship regulations, companies are generally responsible for curating their own censorship lists and that lists used by any two companies will, on average, have little overlap. However, due to the onerous task of compiling these lists, which may contain tens of thousands of keywords or more, companies are often reluctant to invest the resources required to develop their own lists, instead opting to use whatever lists that might be most easily available. Software developers

have been known to take censorship lists with them when leaving companies and to later use them in new products. Furthermore, when comparing a list across a database consisting of as many as thousands of other previously discovered Chinese censorship lists, it can be possible to find one or more lists from which the list in question may have been derived (or lists which may have been derived from the list in question) due to an amount of overlap unexplainable by chance. Therefore, testing from a large sample of other products' lists can be an effective way to find what another product is censoring.

As such, in our second experiment, we sample tested by drawing from a database of thousands of Chinese censorship lists previously discovered on other platforms consisting in aggregate of 505,903 unique keywords. Instead of testing keywords individually, we treated the entire database as a large text by concatenating the unique keywords together ordered by frequency and secondarily lexicographically. By treating the database as a single, large text, we were able to test more content at once, limited only by each search platform's limitations on query length, decreasing the time required to test and increasing the chance of discovering keyword combinations consisting of more than one keyword. When we discover a censored string of text, we isolate its triggering keyword combination and record it. We then resumed testing at the character after the censored keyword combination's earliest character (i.e., after the keyword combination's earliest keyword's earliest character).

Experiment 3: Ongoing testing from news articles

In our third experiment, we test from news articles in a perpetual, ongoing fashion. Our motivation for choosing news articles is that they are easy to collect, contain words related to current events, and often cover politically sensitive topics. Furthermore, they may be directly the desired object of inquiry on a web search platform or the object of discussion on a social media network, as we found many titles of or long phrases in news articles censored on search platforms.

To facilitate this experiment, every 60 seconds, we check for and collect news articles from 16 different RSS feeds spanning Mandarin, Cantonese, Tibetan, and Uyghur languages as well as editorial stances which range from expressly pro-Beijing to expressly critical of Beijing including those news sources with stances in between (see Table 9 for a complete list).

Source	Language
Botan	Mandarin
Boxun	Mandarin
China Digital Times	Mandarin
Deutsche Welle Chinese	Mandarin
Financial Times Chinese	Mandarin
Mingpao	Cantonese
New York Times Chinese	Mandarin
People's Daily	Mandarin

Source	Language
Radio France Internationale Chinese	Mandarin
Radio Free Asia Cantonese	Cantonese
Radio Free Asia Mandarin	Mandarin
Radio Free Asia Tibet	Tibetan
Radio Free Asia Uyghur	Uyghur
Solidot	Mandarin
Voice of America Chinese	Mandarin
Voice of America Cantonese	Cantonese

Table 9: RSS news sources used in testing.

For the purposes of testing, we consider each article text a concatenation of its RSS title, description, and URL. On each search platform, we then test each article as much at a time as possible, as limited by the platform's maximum query length. As in Experiment 2, when we discover a censored string of text, we isolate its triggering keyword combination and record it, and we then resume testing at the character after the censored keyword combination's earliest character.

Experimental setup

We coded an implementation of our experiments in Python using the Selenium Web browser automation framework and executed the code on Ubuntu Linux machines. We tested each search platform from a Toronto network except for Bing, which we tested from a Chinese vantage point using a popular VPN service. Experiment 1 was performed in October 2022. Experiment 2 was performed in February 2023. Experiment 3 began January 1, 2023, and is ongoing as of the time of this writing.

Results

In this section we detail the results of our first two experiments and present preliminary results from our third.

Experiment 1: Censorship of people's names

Among the 18,863 names tested from Wikipedia, we found a combined 1,054 unique names — over 1 in 18 — censored across the search platforms which we tested. Among the unique censored names, 605 were hard censored on at least one platform, and 449 were only ever observed to be soft censored. Among platforms which performed both hard and soft censorship, such platforms performed very little hard censorship, suggesting that they prefer to perform soft censorship when operators possess the capability. From the censors'

perspective, soft censorship may be more desirable as the way in which it controls information is less obvious, but, from the platform operators' perspective, it may be also desirable, as it creates less friction during a user's interaction with the platform because a user, if receiving no results on one platform, may be tempted to try switching to another.

Among the platforms analyzed, we found Weibo to target the highest number of names (474) for some type of censorship, and among the web search engines, we found similar levels of censorship, with Sogou targeting 282, Baidu targeting 219, and Bing targeting 189 with some type of censorship. Strictly concerning hard censorship, web search engines targeted very few names. Baidu hard censored "习明泽" [Xi Mingze], Xi Jinping's daughter, and "徐晓冬" [Xu Xiaodong], a mixed martial artist with anti-China political views. Seemingly beyond coincidence, Sogou hard censored the same two names, although Sogou targeted Xi Mingze with a more broad rule: "习 + 明泽" [Xi + Mingze]. These similar findings are especially surprising as Xu Xiaodong, while a sensitive name, would not seem as sensitive or as well known a name as many others in Chinese politics. We did not find Bing to hard censor any names.

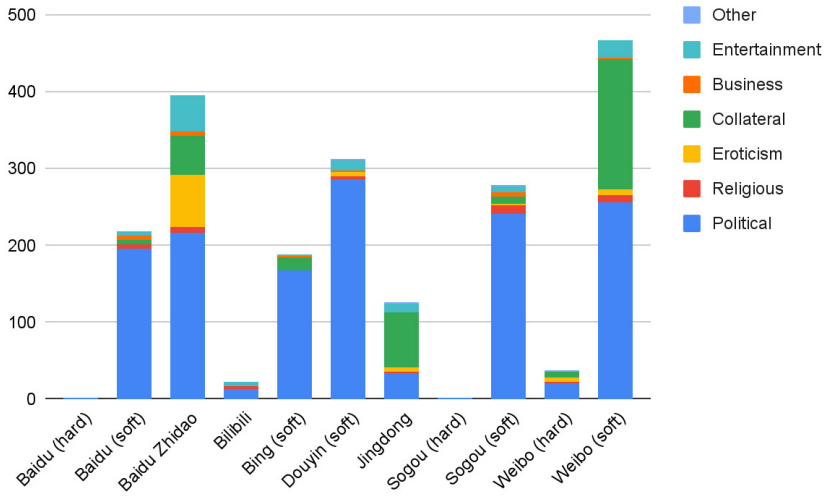


Figure 5: For each platform, for hard and (if applicable) soft censorship, a breakdown by category of the number of names censored in that category.

To better understand the motivations behind search platforms' censorship of people's names, we developed a codebook to categorize each censored keyword according to a person's significance, particularly in the context of Chinese political censorship. Following grounded theory, we first went through all censored names to discern broad categories and repeated themes. This iteration led to seven high-level themes for the codebook. We then reviewed all of the censored names again and applied an appropriate label to each keyword (see Figure 5).

We categorized sensitive names into seven common themes: "Political" (e.g., political leaders and dissidents, major historical events, criticism of the Communist Party, or proscribed political ideas), "Religious" (e.g., banned religious faiths, spiritual leaders, and

religious organizations), “Eroticism” (e.g., pornographic material, adult film actors, sex acts, adult websites, and paid sexual services), “Collateral” (names collaterally censored by a censorship rule targeting someone or something else), “Business” (businesspeople who do not have a clear political motivation for their censorship), “Entertainment” (celebrities, artists, singers and related figures in the entertainment and associated industries who do not have a clear political motivation for their censorship), and “Other” (a residual category that contains content which either does not fit within the other six categories and terms which have been censored for unclear reasons).

We found that most names that platforms censored were for political motivations, whether to shield leaders and other pro-Chinese-Communist-Party members from criticism or to silence dissidents. However, we also found that many names were collaterally censored by rules clearly targeting content other than that person or their name. As examples in English, Hong Kong musical artist “DoughBoy” was soft censored on Weibo for containing “ghB”, GHB being a drug illegal in China and broadly elsewhere, and Baidu Zhidao censored South Korean band “FLAVOR” for containing “AV”, an abbreviation for adult video. As examples in Chinese, Polish Violinist “亨里克维尼亚夫” [Henryk Wieniawski] was soft censored on Weibo for containing “维尼” [Winnie (the Pooh)], a common mocking reference to Xi Jinping, and Microsoft Bing soft censored Chinese actress “习雪” [Xi Xue] for containing “习” [Xi], Xi Jinping’s surname. Weibo’s soft censorship collaterally affected the largest number of names due to the platform’s use of broad censorship rules. In second and third are Jingdong and Baidu Zhidao, respectively. These examples of collateral censorship speak to the methodological importance of not just testing whether a string is censored but also of understanding the exact censorship rule targeting its censorship to avoid misattributing the censor’s motives.

Comparing the soft censorship of social networks Douyin and Weibo, we found that they censor a similar number of names under political motivation, with Douyin censoring slightly more. However, due to its use of broader rules, Weibo had much more names censored collaterally, whereas Douyin’s more specific rules were able to pinpoint political names without collaterally affecting any others in our measurements.

Experiment 2: Censorship of known sensitive content

Among our testing spanning 505,903 unique, previously discovered censored keywords, we found 60,774 unique keyword combinations censored across all search platforms which we investigated. Due to Douyin aggressively fingerprinting and banning our testing, we were unable to complete this experiment for Douyin. We also omit Bing hard censorship results from our discussion in this section as we only discovered four keyword

combinations hard censored by Bing, and we believe that these keyword combinations were measurement artifacts of attempting to measure keyword combination censorship of a machine learning classifier trained to detect pornographic queries (we will discuss this more in the section "Evaluation of our model" later below).

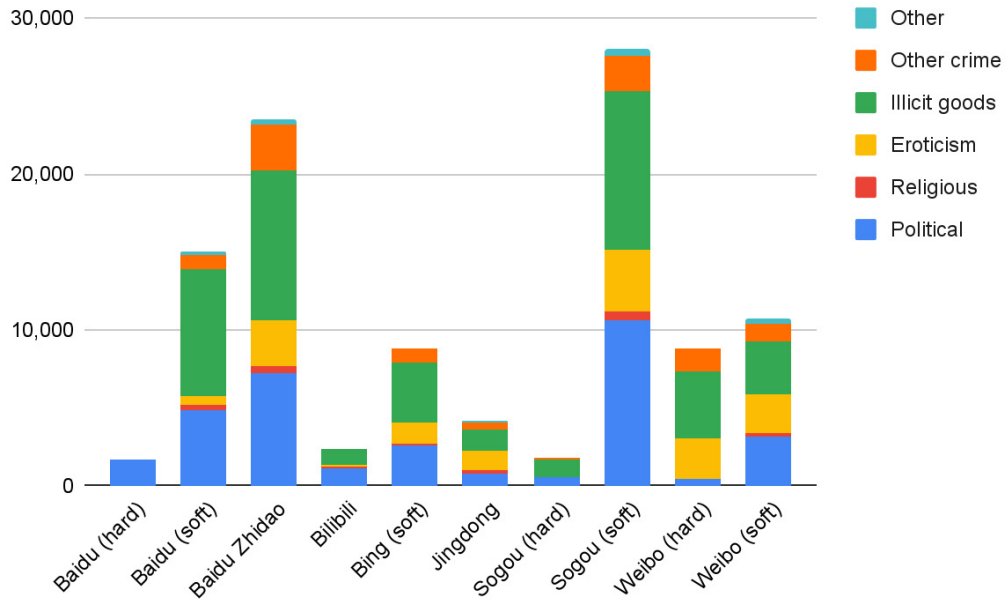


Figure 6: For each platform, for hard and (if applicable) soft censorship, a breakdown by category of the estimated number of keyword combinations discovered in that category.

To understand the type of content censored on each platform, we randomly sampled 200 keyword combinations censored on each platform and categorized them as we did for the previous experiment into themes which resemble but are not all the same as the ones in the previous experiment: "Political" (e.g., political leaders and dissidents, major historical events, criticism of the Communist Party, or proscribed political ideas), "Religious" (e.g., banned religious faiths, spiritual leaders, and religious organizations), "Eroticism" (e.g., pornographic material, adult film actors, sex acts, adult websites, and paid sexual services), "Illicit Goods" (e.g., narcotics, weapons, and chemicals), "Other Crime" (e.g., gambling, fraud, extortion, counterfeiting, and private surveillance), and "Other" (a residual category that contains content which either does not fit within the other five categories and terms which have been censored for unclear reasons). For each platform, based on the proportion of keyword combinations that we found in each category in our random sample, we then estimated the total number of keyword combinations in each category for each platform.

We based our criteria for these six categories on what we found through examining censored content on the eight platforms listed in Table 3. Our six categories also roughly match the categories of prohibited content listed in Chinese government legislation (Figure 1), the targets of internet purification special operations (Figure 2), official announcements on

reports of undesirable or harmful online information (Figure 2), and the nine categories of illegal, undesirable, or harmful information listed on the Cyberspace Administration of China's Reporting Center (Figure 3). Below we describe our findings from each of these categories in more detail.

Political

We found that a large proportion of censored names of political leaders refer to Xi Jinping's name "习近平" or his family. Examples include his current wife "彭丽媛" [Peng Liyuan], his former wife "柯玲玲" [Ke Lingling], his sister "齐桥桥" [Qi Qiaoqiao], and his daughter "习明泽" [Xi Mingze] (see Table 10 for a breakdown per platform).

Baidu (hard)	Baidu (soft)	Baidu Zhidao	Bilibili	Bing (Soft)	Jingdong	Sogou (hard)	Sogou (soft)	Weibo (hard)	Weibo (soft)
1,332	390	510	282	64	53	91	1,754	0	33

Table 10: For each platform, a breakdown of the estimated number of keyword combinations discovered related to Xi Jinping or his family based on sample testing.

Censored terms referring to Xi Jinping included the hard censoring of numerous homoglyphs (e.g., "刁斤干" on Baidu) of Xi's name, as well as hard censoring of terms like "xi + 包子" [xi + bun] on Bilibili, a reference to earlier propaganda campaigns which painted Xi as an avuncular figure with simple tastes. Other references to Xi are soft censored, including some homonyms (e.g., "吸精瓶") and the term "三连任" [three consecutive terms] on Bing, a reference to Xi's third term as China's paramount leader. References to Xi's personal life are also widely censored. These include terms like "离婚 + 习近" [divorce + Xi Jin], possibly referring to Xi's first marriage to Ke Lingling. References to Xi Jinping's daughter Xi Mingze, such as "明泽 + 公主" [Mingze + princess] on Sogou, are hard censored. Little information is publicly available about Xi Mingze, but she is believed to have enrolled in Harvard University in 2010 under a pseudonym. Terms related to other Xi family members are also censored, including references to rumors that Xi Jinping's elder sister Qi Qiaoqiao has Canadian citizenship (e.g., "加拿大籍 + 习近平 + 大姐" [Canadian nationality + Xi Jinping + eldest sister]) which are hard censored on Baidu.

While references to Xi Jinping are the most widely censored among all of China's political leaders, references to other past and current political figures are also censored. Some references to former premier "温家宝" [Wen Jiabao], including homonyms of his name (e.g., "温加煲" on Bilibili), are soft censored, as are phrases like "温 + 贪污" [Wen + corruption] on Weibo, the latter referring to accounts of alleged Wen family corruption covered in an investigative report by The New York Times.

Terms indicating criticism of the Communist Party were also subjected to censorship. These include homonyms for Communist Party (e.g., "共抢党", soft censored on Weibo), as

well as calls for the Communist Party to step down (e.g., "GCD + 下台" [GCD + step down] on Baidu). Some hard-censored slogans, like "洗脑班" [brainwashing class] on Jingdong and "共产党灭亡" [death of the Communist Party] on Bilibili, are associated with material produced by the Falun Gong spiritual movement. For example, the term "退党保平安" [quit the Party to stay safe and peaceful], hard censored on Bilibili, refers to a campaign launched by the Falun Gong to encourage Chinese citizens to quit the Communist Party, Communist Youth League, and the Young Pioneers. Other censored terms refer to the 1989 Tiananmen Square protests and subsequent massacre (e.g., "TAM学生" [TAM students], soft censored on Bing) and notable dissidents (e.g., "晓波刘" [Xiaobo Liu], soft censored on Bing) and "吾尔开希" [Wu'er Kaixi], hard censored on Bilibili).

Religious

Much of the censored content concerning religion refers to banned spiritual groups, in particular the Falun Gong. These include homonyms for Falun Gong (e.g., "法仑功") and references to the persecution of Falun Gong devotees (e.g., "弟子 + 迫害 + 洗脑" [disciple + persecution + brainwash]), both soft censored on Baidu. References to other banned spiritual groups are also soft censored, like "觀音法門" [Guanyin Famen, in traditional characters] on Sogou and "观音法门" [Guanyin Famen, in simplified characters] on Bing and "狄玉明" [Di Yuming], the spiritual leader of Bodhi Gong.

Not all censored religious terms refer to banned spiritual groups. The title of Tibet's exiled spiritual leader, "达赖喇嘛" [Dalai Lama], is hard censored on Jingdong. Terms related to Christianity are also hard censored, though for reasons which are not immediately clear. "耶稣 + 少儿" [Jesus + children], "青少年 + 上帝" [youths + God], and "青少年 + 基督教夏令营" [youths + Christian summer camp] are all hard censored on Jingdong. While authorities have not banned Catholicism and Protestantism, Christian religious activities are strictly monitored throughout China. Authorities also routinely surveil, harass, and detain practitioners of underground house churches and Christian-influenced banned faiths like Church of Almighty God ("全能神教会"). The hard censoring of references to youths and Christianity may also be in response to reported state efforts to prevent those under the age of 18 from participating in religious education.

Eroticism

Censored terms in these categories refer to various kinds of pornographic material, acts or body parts, and paid sexual services. This includes terms like "色情无码" [uncensored pornography], soft censored on Bing, Japanese adult film actor "唯川纯" [Jun Yuikawa], hard censored on Baidu Zhidao, and sex acts like "舔嫩逼" [lick tender pussy], hard censored on Bilibili. Other censored terms refer to soliciting sex workers, such as "包夜 + 按摩" [overnight + massage], soft censored on Baidu, or "婊子上门" [visiting prostitutes], hard censored on Baidu Zhidao, or specific body parts, like "大屌" [big dick], soft censored on Sogou.

Illicit Goods

Many censored terms concerning illicit goods refer to drugs. Some refer to selling drugs like "卖 + 咖啡因" [sell + heroin] or "售 + 摇头丸" [sale + ecstasy], both hard censored on Bilibili, or "售 + 地西泮" [sale + diazepam], soft censored on Baidu. Others terms concern manufacturing drugs such as "制作 + 毒药" [crafting + poison], soft censored on Sogou, or "提炼 + 三甲氧基安非他明" [refining + Trimethoxyamphetamine], hard censored on Bilibili.

Censored terms also refer to weapons, including euphemistic references to particular weapons (e.g., "气狗" [air dog] or air gun, hard censored on Jingdong), their sale (e.g., "批发 + 弓弩" [wholesale + bow and crossbow], hard censored on Weibo), or their manufacturing (e.g., "制作 + 枪" [crafting + gun], hard censored on Weibo).

Chemicals also feature as censored terms, such as "光气 + 提供" [carbonyl chloride + supply], soft censored on Baidu, and references to buying particular kinds of insecticide (e.g., "敌杀磷 + 购买" [Dioxathion + buy], soft censored on Sogou). It is unclear why references to particular chemicals have been censored, though in some cases censorship may be related to the potential use of some chemicals in the manufacturing of narcotics or the production of explosives.

Gambling terms also make up a large number of censored terms related to illicit goods. These gambling-related terms include the names of particular websites (e.g., "金沙sands 线上娱乐场" [Golden Sand sands online resort], a reference to Sands Casino in Macao, hard censored on Sogou), particular kinds of gambling (e.g., "赌马 + 开户" [horse betting + open account], and even "online casinos" in general ("网上赌场", soft censored on Weibo).

Other Crime

This category of prohibited content contains references to a range of illicit or criminal activity. Some refer to various forms of fraud or forgery, including selling high quality counterfeit identity cards ("卖高仿身份证", soft censored on Bing) or searches for police uniforms ("警服", hard censored on Jingdong).

Other censored terms include references to adopting babies ("收养 + 宝宝" [adoption + baby], soft censored on Sogou) to selling organs ("售 + 肝臟" [sell + liver], soft censored on Weibo), potentially censored due to police efforts to deal with child trafficking and kidnapping and illegal organ harvesting, respectively. References to other illicit activities like live broadcasting suicide ("自杀 + 直播" [suicide + live], soft censored on Sogou), hiring kidnapping services ("替人绑架" [kidnap for someone], hard censored on Weibo), or selling diplomas ("卖 + 文凭" [sell + diploma], soft censored on Baidu) are also censored, as are terms related to buying commercial surveillance devices (e.g., "供应 + 卧底窃听软件" [sale + undercover eavesdropping software], soft censored on Baidu).

Other

This residual category included censored terms which did not clearly fit within the other five categories. Some of these were profanity, both in Chinese (e.g., "++你妈" [fuck your mom], hard censored on Jingdong) and English (e.g., "Fucker", soft censored on Weibo). Others referred to news websites blocked in China, like Radio Free Asia ("自由亚洲电台", hard censored on Bilibili) and the Taiwanese newspaper Liberty Times ("自由时报", soft censored on Sogou).

References to censorship itself and how to circumvent content management controls were also censored. This includes references to the Great Firewall ("防火长城", soft censored on Sogou), "翻牆" [leaping over the wall], soft censored on Weibo, and "网络发言防和谐" [online speech anti-censorship], soft censored on Baidu. In other cases, censorship reflected the corporate concerns of a specific platform. Jingdong hard censored the term "狗东" [Dog East or "gou dong"], a satirical play on words referring both to the companies name Jingdong and the use of a cartoon dog as the company's mascot.

Impact of censorship across platforms

Although measuring the number of censorship rules targeting a type of content may be a valuable measure of the amount of attention or resources that a platform has invested into censoring that content, it may be a misleading measure of the actual impact of that censorship. For instance, when looking at Baidu's soft censorship rules, we found 559 keyboard combinations containing the character "习" [Xi]. Many of these are homonyms of Xi Jinping's name (e.g., "习进瓶") or derogatory references (e.g., "习baozi"). Although Baidu uses a large number of rules containing "习", Bing has only has one such soft censorship rule containing "习", but it is to simply censor all queries containing the character "习" without any additional specificity. From this, a naive analysis might conclude that Baidu's censorship of Xi is 559 times broader than Bing's since it has 559 times as many rules, but yet Bing's single, broad rule censors more Xi-related queries than Baidu's long list of specific queries.

To attempt to measure which search platforms had the broadest censorship, we devised a new metric. At first, as an attempt to approximate the number of queries a keyword combination is censoring, we considered using search engine trends data, but such data appeared to have two major issues: first, trends data appeared to have data for only the most common of queries, and, second, trends data for a query were only for queries which exactly matched as opposed to performing a substring match. For example, trends data for "习近" [Xi Jin] would show fewer results than for "习近平" [Xi Jinping], despite "习近" being a substring of "习近平". Thus, to approximate how many queries were censored by a rule censoring all queries containing "习近", we would have to anticipate all such queries that one might make which contain "习近" and then add up the trends data for each.

Instead, we adopted a different metric, which we call the *impact score*, which was devised to approximate the number of web pages censored by a keyword combination. To determine the impact score for a keyword combination rule, we created a query where each keyword in that keyword combination was surrounded by quotation marks. For instance, for the keyword combination "习 + 二婚", we recorded the number of results for the following search query:

"习" + "二婚"

This query requests all pages containing both the exact phrase "习" and the exact phrase "二婚", which mirrors the corresponding keyword combination censorship rule which censors any query containing those exact phrases. For this testing, to obtain the number of web pages impacted by a keyword combination, we measured using Bing as accessed from the Canadian region, as, to the best of our knowledge, Bing has not implemented Chinese political censorship of search results in this region. Note that we apply this metric even to search platforms which are not web search engines, even though these platforms are not searching web pages but rather other items such as store products, microblog posts, or social media videos, as we suspect that this metric can still approximate the impact of the censorship on these platforms as well.

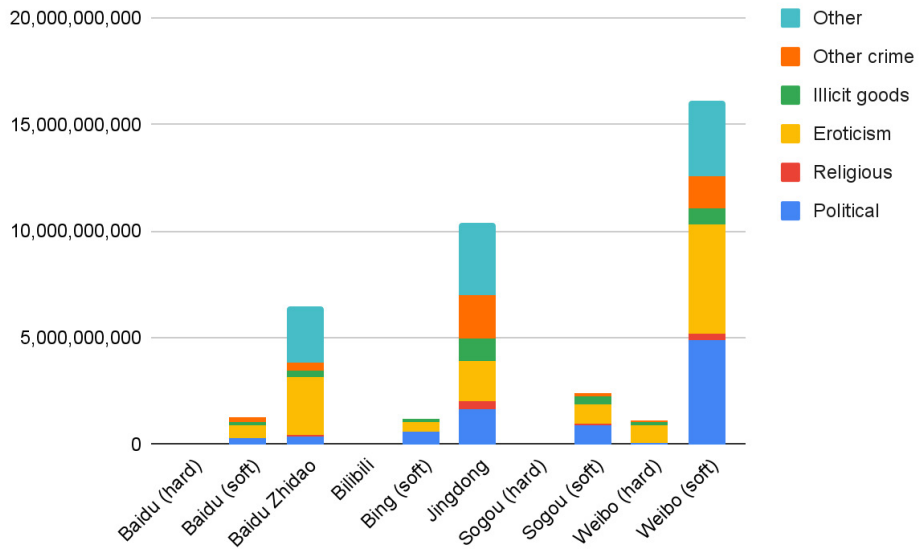


Figure 7: For each platform, for hard and (if applicable) soft censorship, a breakdown by category of the estimated sum of the impact scores of each keyword combination in that category.

To understand the type of content most likely to be censored on each platform, we randomly sampled 200 keyword combinations from each platform by performing a weighted uniform sampling, with replacement, weighted by the impact score of each keyword combination. We then categorized these keyword combinations using the same codebook as before. These results are characteristically different than before, with Weibo

now demonstrating the highest level of total censorship among the search platforms we analyzed (see Figure 7).

Analyzing by category, we find that Jingdong has the highest level of censorship of illicit goods. This finding is not unexpected given that Jingdong is the only e-commerce platform that we analyzed and thus could be expected to have broader filtering of illegal goods.

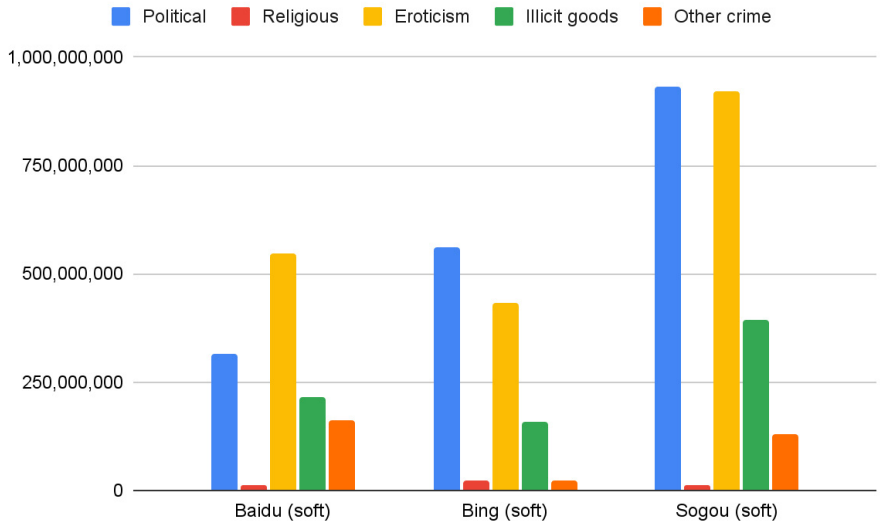


Figure 8: Among web search engines, a breakdown by category of the estimated sum of the impact scores of each soft-censored keyword combination in that category.

Turning our focus to the three web search engines, we find that Sogou has the highest level of overall censorship. Compared to Baidu, Bing has slightly less overall censorship than Baidu. However, breaking down by category, Bing's level of censorship of political and religious topics exceeds Baidu's, with Baidu's filtering of content related to eroticism, illicit goods, and other crimes exceeding Bing's. This finding suggests that Bing is not a suitable alternative to Baidu for users attempting to freely access political or religious content and that to access such content Baidu may be a better choice despite it being operated by a Chinese company.

Experiment 3: Ongoing testing from news articles

In this section we briefly discuss preliminary results from our ongoing experiment measuring censorship rules by testing news articles.

Baidu	Baidu Zhidao	Bilibili	Bing	Jingdong	Sogou	Weibo
1,493	1,426	165	115	329	4,438	908

Table 11: For each platform, as of April 2, the number of new censored keyword combinations which we discovered outside of the previous two experiments.

As of April 2, 2023, since our testing which began January 1, 2023, we have discovered between 155 and 4,438 new keyword combinations on each platform analyzed.

Unfortunately we have limited ability to know if a newly discovered keyword combination was recently added or if it had merely been recently discovered. However, while many of the newly discovered censorship rules could not be shown to be recently added, many others referenced events that occurred since January 1, 2023, seemingly requiring them to have been introduced since then.

As examples, Weibo soft-censored "中国间谍气球" [Chinese spy balloon], referring to a Chinese balloon shot down on February 4, 2023, over the United States which the United States and Canada accused of being used for surveillance, as well as "阮晓寰" [Ruan Xiaohuan], an online dissident who was recently convicted of inciting subversion to state power. Baidu hard censored "逮捕令 + 普京 + 习近平" [Arrest Warrant + Putin + Xi Jinping], referring to an arrest warrant issued on March 17, 2023, by the International Criminal Court for Vladimir Putin. In the days following the issuance, Xi Jinping would visit Putin in Russia. Sogou's soft censorship of the Ukraine crisis used a large number of very specific keyword combinations, many of them referencing 2023 developments:

- 乌克兰 + 王吉贤 [Ukraine + Jixian Wang]
- 博明驳斥美国 + 台湾乌克兰化谬论 [Bo Ming refutes the US + fallacy of Ukrainization of Taiwan]
- 入侵乌克兰一年后 + 俄罗斯依赖中国 [A Year After Invading Ukraine + Russia Depends on China]
- 成为下一个乌克兰 + 台湾 [Be the next Ukraine + Taiwan]
- 王吉贤 + 乌克兰 [Jixian Wang + Ukraine]
- 抗议 + 俄罗斯 + 乌克兰战争 [Protests + Russia + Ukraine War]
- 大疆 + 无人机 + 乌克兰 [DJI + Drones + Ukraine]
- 马斯克 + 乌克兰 + 星链 [Musk + Ukraine + Starlink]
- 俄罗斯 + 入侵 + 乌克兰 + 一年 [Russia + invasion + Ukraine + year]

As we have previously mentioned, our isolation algorithm generalizes effortlessly to all languages. For example, we found that many platforms censored keyword combinations containing Uyghur script. Here are two examples of Bing targeting Uyghur users referring to issues of Xinjiang independence:

- ئەركىنلىك [Freedom]
- ۋەتىنىمىز [Our homeland]

This experiment has only recently begun, and we intend to continue performing this ongoing experiment, measuring how censorship unfolds across these platforms in realtime in response to world events.

Evaluation of our Model

We now reflect on how well our modeling of search platforms' censorship rules as a "list of ordered sequences" fits with their censorship behavior in practice. In general, we found our results to be highly internally consistent using our model. However, both Jingdong and Bing showed inconsistencies which should not be strictly possible in our model. For instance, on Jingdong, we found that both the strings "枪" and "射网枪" are censored even though the string "网枪", which contains "枪", is not. On Bing, we found that both the strings "89天安门" and "1989天安门" are soft censored, even though the string "989天安门", which contains "89天安门", is not. On these platforms, characters surrounding censored keywords can sometimes seemingly play a role in determining whether to censor a string, behavior which is currently not captured by our model, and thus the number of censored keyword combinations may be underreported on these platforms.

While these were minor departures from our model, a more extreme case would be all four keyword combinations which our algorithm found to be hard censored on Bing:

- 台湾 + 小穴 + 护士做爱 + 台湾 [Taiwan + pussy + nurse sex + Taiwan]
- 片BT下载BANNED骚逼 [Piece BitTorrent Download BANNED Pussy]
- 你 + 你 + 你 + 你的屄 [you + you + you + your cunt]
- 你 + 你 + 你的屄 + 你 [you + you + your cunt + you]

Unlike our results for other platforms, including those of Bing's soft censorship, the results for Bing's hard censorship are, while seemingly related to eroticism, mostly nonsensical to human interpretation. The results pages for each of these four queries showed an explicit notification that results were blocked due to a mandatory "safe search" filter being applied to the mainland Chinese region, and we suspect that we were triggering a machine learning classification system trained to detect search queries related to eroticism. While machine learning algorithms struggle to censor according to subtle, broad, and rapidly evolving political criteria, they are more effective at detecting relatively narrower, more well-defined, and more slowly changing criteria such as whether a query is related to pornography. As such, these results may be an interesting glimpse into what would happen if we applied our isolation algorithm against a censorship system applying a machine learning classifier intending to politically censor content.

Authorized domain lists

All three web search engines which we analyzed performed soft censorship, a censorship scheme in which if a query contained a soft-censored combination of keywords, then results would only be returned from a list of authorized domains. In this section, we explore whether different search engines authorized different domains and whether different domains are authorized for different keyword combinations.

To investigate these questions, first we developed a method to measure whether a domain was authorized to be displayed in results for a given string. Our method is a simple modification of the one which we used to determine whether a string is soft censored in general: we replace “site:microsoft.com”, which was a domain which we presumed would not be authorized for any soft-censored string, with “site:IsThisAuthorized.com”, where IsThisAuthorized.com is a domain which we wish to test to see if it is authorized for that soft-censored string. Using this method, we tested across a set of domains *D* and a set of strings *S*.

To choose *S*, we selected those name test strings which we found soft censored on all three web search engines in Experiment 1. To determine *D*, for each of those strings, on each platform, we then searched for these strings, recording all domains which we observed in the first 100 search results. *D* then is the set of all domains which we observed during this procedure. In our experiment, *S* consisted of 83 strings, and *D* consisted of 326 domains.

In October 2022, on each web search engine, for each domain in *D* and string in *S*, we tested whether that domain was authorized for that string on that search engine. We then collected the results into a two dimensional matrix. To draw out the general shape of the lists, we hierarchically clustered both dimensions of the matrix according to the UPGMA method.

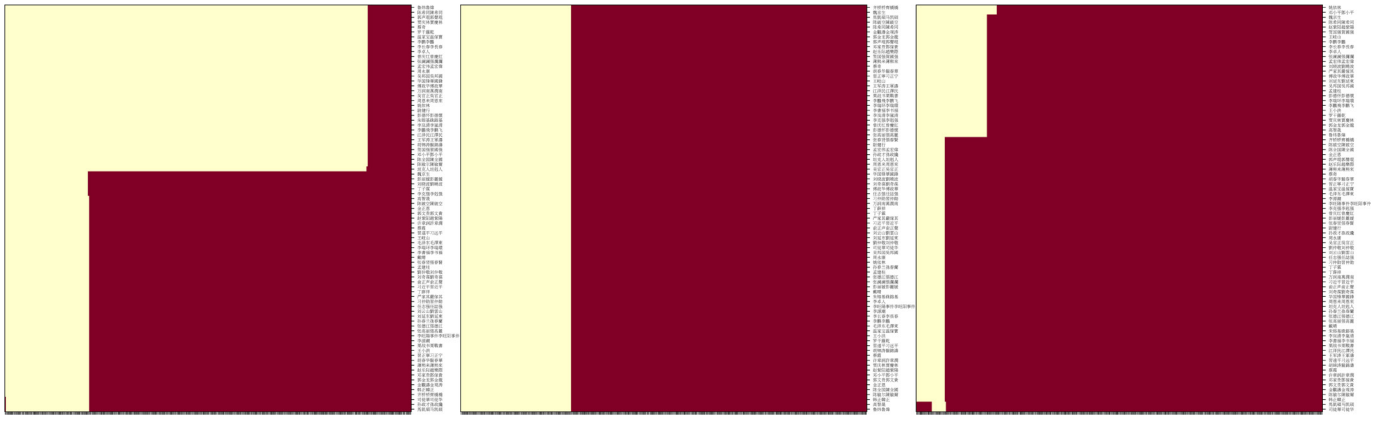


Figure 9: The “shape” of the authorized domains lists for Baidu (left), Bing (center), and Sogou (right): for each domain (x axis) and string (y axis), whether the domain is authorized for that string (light yellow) or not (dark red).

We found disparate authorization lists across web search engines (see Figure 9). We found that, in our experiment, Sogou authorized, on average, the fewest domains for each string, followed by Bing, with Baidu authorizing the most. Bing used the same authorization list for each string which we tested, whereas Baidu appeared to use approximately two different lists, although some strings used lists with small additions or subtractions from these two. Sogou appeared to mostly use two lists, with a third and fourth list being applied to some tested strings. In comparing Baidu and Bing, Baidu had a more complicated set of authorizations, whereas Bing broadly applied the same list to each string and

thus authorizes fewer domains overall. While one might hypothesize that more sensitive keyword combinations are associated with shorter lists of authorized domains, we surprisingly did not notice any correlation between sensitivity and authorized domain list length.

To better understand the domains authorized by these search engines, we categorized them into three categories (see Table 12). The majority of sites on each list were Chinese state-approved news sites. Examples include xinhua.org (Xinhua News Agency), people.cn (People's Daily), and qq.com (QQ News). Sites from this category professed varying degrees of loyalty to the Chinese Communist Party (CCP), ranging from presenting the necessary regulatory license to practice journalism to statements such as these from huyangnet.cn (Authoritative information release platform of Xinjiang production and Construction Corps) indicating Party sponsorship (translated): "Bingtuan Huyang.com is a key news portal website of Bingtuan, which is approved by the Information Office of the State Council and sponsored by the Propaganda Department of the Party Committee of Xinjiang Production and Construction Corps."

List	News	Party-state	Other	Total
D	241 (73.9%)	77 (22.8%)	8 (2.45%)	326
Baidu (short)	45 (64.3%)	24 (34.3%)	1 (1.43%)	70
Baidu (long)	221 (73.4%)	76 (25.2%)	4 (1.33%)	301
Bing	82 (89.1%)	8 (8.70%)	2 (2.17%)	92
Sogou (shortest)	11 (84.6%)	2 (15.4%)	0 (0.00%)	13
Sogou (shorter)	22 (91.7%)	2 (8.33%)	0 (0.00%)	24
Sogou (longer)	50 (84.7%)	8 (13.6%)	1 (1.69%)	59
Sogou (longest)	51 (76.1%)	13 (19.4%)	3 (4.48%)	67

Table 12: Breakdown of authorized domain lists by category.

Other sites were more directly operated by either the Chinese Communist Party or the Chinese state. Many of these were official government web sites of different jurisdictions, such as xinjiang.gov.cn, the official web page of the People's Government of Xinjiang Uyghur Autonomous Region, and gqt.org.cn, the official website of the Chinese Communist Youth League.

Finally, we have a small residual category, which contains miscellaneous sites such as search engines, those providing health information, etc.

One behavior which we were interested in understanding was how search engines behaved when two different soft-censored strings with different authorization lists occurred in the same query. Depending on how the systems are implemented, search engines may prefer the first observed keyword combination (such as if all censorship rules were implemented using a single deterministic finite-state automaton) or they might take the set intersection of all of the authorization lists for each occurring keyword combination. We found,

however, that, when testing two different censored strings with different authorization lists, the list of one is preferred over the other regardless of their positions with respect to each other in the query (see Tables 13 and 14). This finding is consistent with a system which iterates over a list of blocked keyword combinations, testing for their presence in a query, and where, as soon as one is found present, the corresponding action for that keyword combination is taken, aborting the rest of the search.

	baidu.com	mofcom.gov.cn
李克强李剋強	Authorized	Unauthorized
司徒華司徒华	Unauthorized	Authorized
李克强李剋強司徒華司徒华	Authorized	Unauthorized
司徒華司徒华李克强李剋強	Authorized	Unauthorized

Table 13: On Baidu, when both "李克强李剋強" [Li Keqiang, in both simplified characters and traditional characters] and "司徒華司徒华" [Situ Hua, in both traditional and simplified characters] are present, the authorized domains list for "李克强李剋強" is used.

	tibet.cn	baidu.com
韩正韓正	Authorized	Unauthorized
馬凱碩马凯硕	Unauthorized	Authorized
韩正韓正馬凱碩马凯硕	Authorized	Unauthorized
馬凱碩马凯硕韩正韓正	Authorized	Unauthorized

Table 14: On Sogou, when both "韩正韓正" [Han Zheng, in both simplified and traditional characters] and "馬凱碩马凯硕" [Ma Kaishuo, in both traditional characters and simplified characters] are present, the authorized domains list for "韩正韓正" is used.

While this may seem like a mundane finding, it suggests that the original order of keyword combinations discovered on the list can be reconstructed, at least partially, as the order of two keyword combinations with the same authorization lists cannot be directly compared using this method. Such an information side channel could be useful in measuring when a keyword combination was added to the list, where otherwise we would only know when it was first discovered on the list. Furthermore, not just knowing which keyword combinations are censored but also their order on a blocklist can be helpful for inferring how such lists of censorship rules are shared among companies, developers, and other actors, as two lists might have many censorship rules in common by coincidence but, as the number of common censorship rules grows, it becomes super-exponentially unlikely that both lists would have those censorship rules in the same order purely by coincidence.

Figure 10: An example geoblocking block page for a popular Chinese news site.

Finally, when we were categorizing the domains, we noticed that a surprisingly large number of sites were inaccessible from outside of China. We found that, among the 338 domains in *D*, when accessed from a Toronto network, 59 (17.5%) failed to return an HTTP 200 status for either that domain or for that domain preceded by "www.". Some sites appeared to block connections on an IP or TCP layer, whereas others presented application layer block pages (see Figure 10). While the motivation for Chinese censors blocking non-Chinese sites from Chinese access is well understood, it is less understood why Chinese sites are in turn blocking access to non-Chinese users. Future research is required to understand this troubling progression of the balkanization of the Internet.

Limitations

Our study analyzes automated censorship of search queries across a variety of platforms. However, there exist other layers of censorship which might also be affecting search results. For instance, on social media platforms, posts may be automatically or manually deleted or shadow-banned if they contain sensitive content. In fact, the rules for automatically censoring posts may often match the rules for censoring search queries (see Appendix A). Users may also self-censor under the fear of reprisal for posting sensitive content. However, our work analyzes the rules used by platforms to automatically censor search results but not any of the other factors which might be skewing those results.

Some search platforms may have some censorship rules which censor not according to whether a query contains certain keywords but whether the query exactly equals a certain keyword. While this may seem like an inflexible manner to censor queries, we have observed such a case on Weibo, specifically when searching by hashtag, when we observed one hashtag which was censored (e.g., #hashtag) but superstrings of that hashtag which were not (e.g., #hashtagXYZ). Our method will often fail to detect censorship rules such as these which require exact matches, as our isolation algorithm requires that any query containing the censored content be censored in order to isolate the content triggering censorship.

Discussion

As North American technology companies such as Google mull over whether to expand search or other services to the Chinese market, a popular argument has been that, although infringing on users' political and religious rights is inherently wrong, perhaps a North American company could better resist Chinese censorship demands and provide a less-infringing service than a Chinese company. However, even if the ends are to justify the means, then, for this argument to have any validity, the service provided by the North American company must be less infringing.

Unfortunately, our study provides a dismal data point concerning this argument. It suggests that whatever longstanding human rights issues pervade in China, they will not be magically addressed by North American technology companies pursuing business in the Chinese market. To the contrary, our report shows that users using Microsoft Bing are subject to broader levels of political and religious censorship than if they had used the service of Bing's chief Chinese competitor. In fact, rather than North American companies having a positive influence on the Chinese market, the Chinese market may be having a negative influence on these companies, as previous work has shown how the Chinese censorship systems designed by Microsoft and Apple have affected users outside of China.

The methods introduced in our work facilitate future, ongoing censorship measurement. In light of our third experiment, we presented preliminary results from an ongoing experiment discovering search platform censorship by sampling text from news articles. We intend to continue running this experiment for the indefinite future, tracking changes to search platform censorship over time as events around the world unfold.

The challenges in moderating search queries are similar to those moderating queries to machine-learning-powered chat bots such as ChatGPT in that, just as with search platforms, when compared to the understanding of the actual query evaluator, the censorship system may have an inconsistent understanding of a query which can be exploited to measure for the presence of censorship. As one possible example, AI researcher Gary Marcus found through experimentation that ChatGPT responded to the query, "What religion will the first Jewish president of the United States be?" as follows: "It is not possible to predict the religion of the first Jewish president of the United States. The U.S. Constitution prohibits religious tests for public office… It's important to respect the diversity of religions and beliefs in the United States and to ensure that all individuals are treated equally and without discrimination." The ChatGPT query he used to reveal its censorship is tautological in a way reminiscent of our use of truisms to reveal search engine censorship. In the same way that tautological queries which should have guaranteed answers can be used to measure chat-bot censorship, our work uses truisms, which

should have guaranteed search results, to measure search platform censorship rules. We hypothesize that, as Baidu or Microsoft, introduce their chat bots into the Chinese market, a similar use of such tautological queries or truisms can be used to flesh out what political censorship rules these bots implement to comply with Chinese political censorship laws and regulations.

Finally, we hypothesize that the methods we used to test for search engine censorship can be adapted to evade censorship as well. In this report, to measure search platforms' censorship rules, we utilized the inconsistency between the censorship filter's and the query parser's understanding of specially crafted queries. In Appendix A, we present a proof of concept demonstrating how this inconsistency can be exploited to evade censorship on Weibo. We leave it to future researchers to exploit this inconsistency to develop additional evasion techniques to evade Weibo search censorship as well as to evade search censorship on other search platforms at large.

Availability

For Experiment 1, the list of names we discovered to be hard or soft censored on each platform as well as the keyword combination triggering their censorship is available here. For Experiment 2, the list of keyword combination rules which we discovered to be triggering hard or soft censorship on each platform tested is available here. For our authorized domains experiment, for each platform, a matrix detailing whether each tested string was authorized or unauthorized for each tested domain is available here. The algorithms which we use to isolate which combination of keywords is triggering censorship of a string are available here.

Appendix A: Evasion of Weibo search censorship

A technique in our report, which makes many of our methods possible, is to exploit how the censorship system has a different, typically more naive understanding of a search query than the search platform proper. For instance, when testing Baidu Zhidao, in order to guarantee a nonzero number of search results in the absence of censorship, we test a query by searching for combining two predicates: the presence of a common word and then for the lack of presence of a sensitive keyword, e.g., "the -(xi jinping)". While an intuitive understanding of this query might suggest that excluding sensitive content should not subject the query to censorship, the censorship filter nevertheless acts on a different level of understanding, by simply scanning the query string for the presence of sensitive keywords. In the remainder of this appendix, we present a proof-of-concept exploitation

of the same gap in understanding to demonstrate a method of evading Weibo's search censorship, which can be used by censorship researchers in determining "ground truth" search results as well as motivating other techniques for evading search censorship on Weibo and platforms at large.

The technique we present is simply to put an underscore (_) between at least one of (or all of) the Chinese characters in a censored keyword. For instance, at the time of this writing, "法轮" [Falun] is hard censored by Weibo search. However, "法_轮" returns results for posts containing "法轮", whereas separating the characters with a space ("法␣轮") merely returns results for posts containing "法" and "轮", although not necessary adjacently. The reason why placing underscores evades censorship is due to the fact that the search platform's query parser seemingly removes underscores between Chinese characters before evaluating the query but the search filter performs no such removal before scanning the search query for sensitive content.

Using underscores in this manner to evade Weibo search censorship appears to have some limitations. First, we have not observed this evasion to work with English letters, and so it only applies to evading censored keyword combinations which contain at least one keyword containing at least two Chinese characters. Moreover, we have not observed this evasion to work when searching by hashtag, as the query parser does not appear to silently remove underscores in such searches. Finally, in addition to search queries, posts themselves on Weibo may also be subject to automated deletion based on the presence of sensitive combinations of keywords. Thus, when searching for a keyword or combination of keywords that is simultaneously banned from appearing in posts, bypassing the search censorship will still yield zero results as there are no matching posts to return as they have already been deleted. While the first two limitations discussed in this paragraph may be overcome by further development of this evasion technique, any method for evading search censorship rules will still be limited to only returning results which were not also victim to content deletion rules.