Measuring QQMail's automated email censorship in China

Jeffrey Knockel
Citizen Lab, University of Toronto
jeff@citizenlab.ca

Lotus Ruan
Citizen Lab, University of Toronto
lotusruan@citizenlab.ca

ABSTRACT

Desiring to control political ideas shared within its borders, China fosters its own domestic Internet communications platforms where it can better enforce its laws and more tightly control political expression. Access to foreign platforms failing to comply with China's requirements on speech restrictions are often wholesale blocked by the country's national firewall. However, some technologies, such as the Web and email, are too universal to wholesale block and are restricted discriminately. While China's restrictions to Web access have been extensively studied, our work provides a look into email censorship in China.

In this work, we study censorship on Tencent's QQMail, the most popular email platform in China. We introduce a technique to test whether QQMail automatically censors a message, without requiring control of any Internet endpoints in China or any accounts on the QQMail platform, finding 173 combinations of keywords that trigger automated censorship of multiple political topics. We also discover that some censored keyword combinations have *extenuating* keyword combinations, any of whose presence disables the corresponding censored combination's censorship. Many of these extenuating terms are as politically sensitive as the censored combinations they are extenuating. We find that the motivation behind such extenuating terms is puzzling, defying easy explanation.

CCS CONCEPTS

• **Networks** → **Network monitoring**; • **Social and professional topics** → *Censoring filters*;

KEYWORDS

China, censorship, email

ACM Reference Format:
Jeffrey Knockel and Lotus Ruan. 2021. Measuring QQMail's automated email censorship in China. In *Workshop on Free and Open Communications on the Internet (FOCI '21), August 27, 2021, Virtual Event, USA.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3473604.3474560

1 INTRODUCTION

Email has long been known to be subject to automated censorship in China [4, 28]. However, despite extensive research of how China censors globally decentralized Internet technologies such as HTTP and DNS, there has been limited work to study and characterize which topics are censored in email in China. In this work, we analyze censorship performed by Tencent's QQMail, the most popular email platform in China [2].

We study how QQMail censors email messages depending on the presence of certain combinations of keywords. We say that a keyword combination such as 习近平 + 恢复帝制 (Xi Jinping + restore the monarchy) is present in a message if and only if each component, 习近平 and 恢复帝制, is present somewhere in the message.

Our work provides the following contributions:

- We introduce a method to measure automated email censorship on QQMail without requiring an account on the platform or an Internet endpoint in China.
- We uncover 173 combinations of keywords that trigger automatic email censorship.
- We discover that some censored combinations have a unique set of keyword combinations, any of whose presence disables their corresponding keyword combination's censorship from triggering.
- Through a side channel in QQMail's censorship system, we find that we can measure whether another address has previously received email from a QQMail user.

2 RELATED WORK

There has been extensive research characterizing Chinese censorship. Some research has focused on how blog [12, 16] or micro-blog [17, 26] posts are chosen to be deleted by moderation teams on domestic Chinese platforms. While human review of persistent media is viable, such review is impractical for censoring Web browsing or real-time communication, which must be done according to automated rules.

Much work has focused on analyzing the automated censorship of decentralized Internet technologies in China. Work analyzing Chinese Web censorship most resembling ours is that characterizing how keywords in DNS queries [1, 11] or HTTP requests [6, 19, 22] trigger censorship from the Great Firewall of China (GFW), China's national firewall. We expand upon such work by characterizing what content sent via email, another decentralized Internet technology, is censored in China.

Fen Qin previously documented how the GFW censors non-encrypted SMTP email traffic according to sender and

FOCI '21, August 27, 2021, Virtual Event, USA

ACM ISBN 978-1-4503-8640-1/21/08.
https://doi.org/10.1145/3473604.3474560

recipient, identifying one such blacklisted sender and recipient [28], and Bock *et al.* [3] subsequently identified server-side methods to evade such SMTP censorship, in addition to other types of network censorship. The subject of our work differs in two ways from this SMTP censorship, in that we measure censorship of message body content and that the censorship which we measure is not performed by the GFW.

As the email censorship we measure is performed by QQMail's email servers, our work is also similar to that studying how private social media companies implement automated censorship, such as that of chat apps [5, 10, 20], live streaming apps [13], and games [15]. Notably, research [20] studying chat censorship on WeChat, another Tencent-operated platform, found that messages are not just censored by the presence of single keywords but by the presence of certain combinations of keywords. Our work studying keyword combination censorship on Tencent's QQMail expands on this by discovering and characterizing extenuating keyword combinations and which censored combinations their presence extenuates.

3 MEASUREMENT DESIGN

In this section, we first explain our method for testing whether a message is censored by QQMail. We then generalize this method to discover combinations of keywords that censor messages as well as each of their extenuating combinations of keywords, which disable those censored combinations' censorship.

3.1 Testing message censorship

In a typical scenario, when a user writes an email and hits "send", that user's email client connects to their email provider's SMTP (simple mail transfer protocol) server, communicating the email to it using the SMTP protocol. Their provider's SMTP server then uses DNS to look up the MX (mail exchange) server corresponding to the domain in the recipient's email address, using the SMTP protocol again to relay the email to that MX server. The MX server then delivers the email locally. In our testing, we behave as the user's email provider's SMTP server would in the above scenario and connect directly to a QQMail MX server to send email messages.

To determine whether a message is censored, we analyze the response of the MX server to sending the message with the SMTP DATA command. We found that when sending emails containing a censored message, the MX server responds to the DATA command with response code 550 and line "Mail content denied", whereas without censored content the server would return code 250 line "OK". However, this method undesirably sends an email to a QQMail user, whose account may be subject to adverse action.

Therefore, we perform this test by using a recipient address that does not exist. When sending a message without censored content to a recipient that does not exist, we found that the server replies to the DATA command with code 500 and line "Mailbox unavailable or access denied". However, if the message contains censored content, it still returns "Mail content denied", even if the recipient does not exist. Thus, we determine whether messages are censored by distinguishing between these two response lines without any email ever being delivered to an end user.

In our testing, we use a sender address with 12 randomly chosen letters for its user name and another 12 randomly chosen letters followed by ".com" for its domain. We chose this address once and fixed it across all testing.

Finally, we encode each message in UTF-8. To help ensure that we are not incidentally measuring the censorship of a network middlebox or interference from the Great Firewall of China, we then base64-encode the message in keeping with RFC 2045. Finally, although we found that QQMail's MX servers do not support TLS, we upgrade the connection to STARTTLS after connecting and before transmitting the message to provide further protection from passive eavesdropping.

3.2 Discovering censored keyword combinations

To discover which combinations of keywords triggered message censorship, we used sample testing. In our testing we discovered that only the first 102,400 UTF-8 bytes of the message were scanned for censored keyword combinations. This limit exists regardless of the encoding of the sent message, suggesting that the censorship system decodes all messages and then re-encodes them to UTF-8 before analyzing them. Thus, in all of our testing, we used this limit as our *maximum testing size.*

We created samples for testing using two different methods. The first method sourced test material from tweets that we collected using the Twitter streaming API, an API to collect tweets in real-time. This API requires a string filter and does not support filtering by CJK characters. This, we filtered tweets by the logical disjunction of the presence of the following strings: t, co, com, cn, net, org, RT, http, https, www. These strings were chosen to broadly and with minimum bias select tweets that might not otherwise contain non-CJK content. We also restricted the results to tweets which the API determined to be of Chinese language. To create an email message test sample, we concatenate collected tweets character-by-character, skipping any character which would cause the message to be censored according to rules which we have already discovered. The sample is complete when it has reached our maximum testing size. We call creating samples in this fashion the *Twitter-sourced* method.

Our second method for creating test samples sources material from previously discovered censored keyword lists. We used previously discovered keywords in previous work analyzing Great Firewall HTTP censorship [6] and censorship in chat apps [5, 10, 20], live streaming apps [13], games [15], and GitHub projects [14]. We concatenated together keywords in a similar fashion as we did tweets to create large messages full of sensitive keywords but that were not known to be censored according to any discovered rule. We call creating test samples in this fashion the *keyword-list-sourced* method.

After creating test samples, we tested whether they were censored using the test method described in the previous section. Since QQMail consistently censored content, if a test sample was censored, we initially used Xiong *et al.*'s *component-aware binary splitting* (CABS) algorithm [25] to isolate the keyword combination triggering its censorship. The result of this algorithm is a minimal set of keywords triggering censorship of the test sample, *e.g.*, 千古一帝 + 习近平 (one emperor + Xi Jinping). However, early in our testing we discovered that, if a message contained a censored keyword combination, it might not be censored if it contained other *extenuating* combinations of keywords whose presence disables the censorship of their corresponding censored combination. We found that, in general, each censored keyword combination had a distinct set of extenuating combinations.

This discovery was significant because, not only is measuring the exceptions to censorship rules important for holistically characterizing the system of information controls, but we found that the existence of extenuating combinations affected the correctness of the CABS algorithm. While any output from the algorithm would trigger censorship, it would not necessarily be a minimal, most general combination, in that it may have spurious components (aa + bb vs. XX + aa + bb) or its components may have spurious characters (aa + bb vs. aaXX + bb) or be unnecessarily combined (aa + bb + cc vs. aabb + cc). These under-generalizations occur when, during isolation of the triggering keyword combination, an extenuating combination is removed from the test sample, activating the censorship of a new keyword combination, causing some of the previously learned criteria about what is required to trigger censorship to be no longer applicable.

To isolate a triggering keyword combination from the possibly under-generalized output of the CABS algorithm, we run the output of the CABS algorithm through CABS again. While it is unlikely, it is possible that another extenuating combination was removed, resulting in the same issue as before. Thus, to ensure correctness, to isolate censored keyword combinations from censored test samples, we iteratively run CABS until the result converges, *i.e.*, until the input and output to CABS are the same. We call this algorithm *iterated CABS*.

We found early in our testing that if we attempted to test samples too frequently, we would be rate limited by the MX server, receiving response code 550 with line "Connection frequency limited" and our IP address would be temporarily banned for approximately 24 hours. We thus limit testing messages to every 30 seconds, a value which we found sufficiently infrequent to trigger rate limiting.

3.3 Discovering extenuating combinations

To discover each censored keyword combination's extenuating combinations, we used a method similar to how we found the censored keyword combinations. To generate a test sample, we begin with a randomly chosen censored keyword combination. Next we complete the sample by concatenating character-by-character randomly chosen keywords using the keyword-list-sourced method described earlier. We skip any character that would introduce another known censored keyword combination or that would introduce a known extenuating combination of the censored keyword combination being tested. The result of this process is a test sample consisting of a known censored keyword combination with other words such that, according to our understanding of that censored keyword combination's extenuating combinations, should be censored.

When such a test sample is found to not be censored, we isolate a responsible extenuating combination using a modified version of the CABS algorithm called *CABSE*. The principle modifications are to always include the censored keyword combination in all tests and, since we are attempting to find the keyword combination responsible for the sample *not* being censored, compared to CABS, we take the opposite branch after every inquiry into whether a test sample is censored (see Algorithms 1 and 2 in Appendix A). Because of an analogous problem to earlier, where in this case another censored keyword combination may appear in the test sample if one of its extenuating combinations is also present, we run an iterated version of CABSE which we call *iterated CABSE*.

When we find a new censored keyword combination, we test the set of all known extenuating combinations across all censored keyword combinations to see which, if any, also extenuate the new combination. Moreover, when we discover an extenuating combination which we had not previously discovered to extenuate any other censored keyword combination, for all other censored combinations, we test to see if it extenuates that censored combination.

We performed our measurement of extenuating combinations from a separate process and IP address.

4 RESULTS

We performed all testing between April 1 and May 9, 2021. We found 173 censored keyword combinations. Each censored keyword combination had between zero and 38 discovered extenuating combinations. Across all censored combinations, we found a total of 43 unique extenuating combinations. All censored keyword combinations consisted entirely of Chinese characters. However, some of their extenuating combinations contained English letters and punctuation.

In the remainder of this section, we report our findings concerning censored keyword combinations and their extenuating combinations.

4.1 Analysis of censored keyword combinations

Informed by the grounded theory approach [8], we reviewed all of the censored keyword combinations that we discovered and developed a codebook based on the common themes that emerged. We then categorized all censored combinations based on that codebook. For each keyword combination, we also conducted further research on the events associated with it to gain a better understanding of its underlying context.

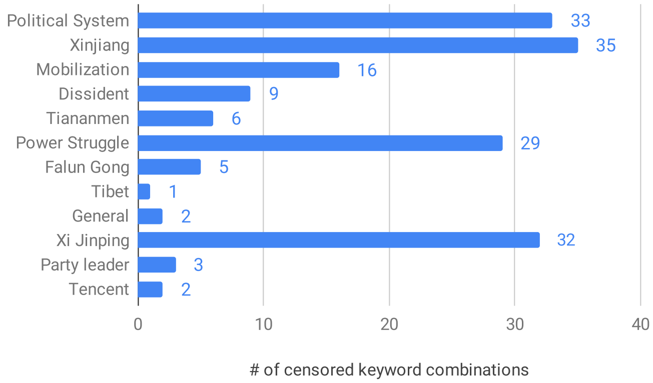

Figure 1: The # of censored keyword combinations in each category

While the keyword lists from which we randomly generated our keyword-list-sourced test samples contained keyword content related to spam, phishing, gambling, pornography, and the sale of illicit goods, none of the discovered censored combinations were related to these topics. Rather, all discovered censored keyword combinations appear politically or commercially motivated. Most censored political content pertains to China's Xinjiang province, followed by those referencing the Chinese Communist Party (CCP)-ruled political system, China's current leader Xi Jinping, and power struggle within the CCP (see Figure 1 for their distribution).

Xinjiang: Thirty-five out of the 173 keyword combinations pertain to the contentious politics surrounding China's Xinjiang region and Uyghur minorities. These keyword combinations predominately referenced the East Turkestan independence movement (*e.g.*, “东突同胞解放”, “the liberation of East Turkestan compatriot”), a political movement that seeks the independence of Xinjiang/East Turkestan from China. While some organizations within the movement are nonmilitant, many have been labelled terrorist organizations by the United Nations and other governments [23]. A few of the keyword combinations appear to be inciting violence among Uyghurs (*e.g.*, “新疆人民 + 发起圣战”, “Xinjiang people + launch jihad”). Notably, besides censoring content critical of the Chinese government's policies towards Uyghur minorities (*e.g.*, “打压维吾尔人”, “suppress Uyghurs”), censorship also targets content that discriminates (*e.g.*, “低劣种族 + 新疆”, “low-class ethnicity + Xinjiang) or stereotypes Uyghurs (*e.g.*, “新疆人都是小偷”, “Xinjiang people are all thieves”).

Political System and **Xi Jinping**: Content referencing China's authoritarian one-party political system and its incumbent president Xi Jinping was targeted at the almost same rate (19% and 18.5%). These keyword combinations are predominately critical or satirical in nature. Much of the Political System-related content also implies a call to action, demanding the resignation of the CCP leaders (*e.g.*, “中共中央领导人集体下台”, “CCP central leaders step down”) or the termination of the one-party system (*e.g.*, “争自由反专制”, “fight for freedom oppose dictatorship”). Most Xi-related keyword combinations concern his removal of the CCP's informal presidential term limits in 2018 [7], which describe the move as China's return to a monarchy (*e.g.*, “复辟 + 习近平”, “restore monarchy + Xi Jinping”). Notably, a few keyword combinations referenced an online parody account mocking Xi Jinping which was first launched around 2017. This suggests that the censorship rules might have received a major update since 2017.

Power Struggle: One of the most commonly targeted content categories is keyword combinations referencing power struggle among party leadership (29 out of 173). Most of these keyword combinations pertain to the CCP's most recent and high-profile intra-party power struggle involving Bo Xilai, former member of the CCP Politburo (*e.g.*, “薄熙来仕途”, “Bo Xilai's political career”). Bo was allegedly conspiring with Zhou Yongkang, former senior leader of the CCP, to stage a coup in 2012 (*e.g.*, “周永康兵变”, “Zhou Yongkang military coup”). Despite the scandal occurring nearly a decade ago, that this content still triggers censorship indicates the high sensitivity of the topic of power struggle in China.

Mobilization, **Dissident**, and **Historical Events**: Other topics that attract political censorship on QQMail include content referencing mass mobilization (*e.g.*, “全国各大城市集合”, “nationwide assembly in major cities”), dissidents (*e.g.*, “释放刘晓波”, “release Liu Xiaobo”), the 1989 Tiananmen Square Movement (*e.g.*, “天赋人权 + 勿忘六四”, “natural rights + don't Forget June 4), Falun Gong and its prominent publications (*e.g.*, “大纪元”, “Epoch Times”), and Tibet issues (*e.g.*, “经达赖”, “through Dalai”).

Commercially Motivated Censorship: We found two keyword combinations negatively referencing Ma Huating, Tencent's founder and CEO (*i.e.*, “婊子 + 马化腾”, “bitch + Ma Huating”; “马化腾 + 人渣”, “Ma Huating + scum”). While small in scale, they suggest that censorship on QQMail, Tencent's product, is partly motivated to protect the reputation of its parent company and owner.

4.2 Analysis of extenuating combinations

Whereas censored content is predominately political, extenuating combinations include an even amount of political, technical, and miscellaneous content. Much of the extenuating content consists of domain names of popular URL shortening services such as those of Google, Baidu, and Weibo (*e.g.*, goo.gl/, dwz.cn/, and t.cn/). Other extenuating combinations are politically sensitive references such as a reference to human rights defender 刘士辉 (Liu Shihui). Controversial religious movement 法輪功 (Falun Gong) is also an extenuating combination, although only in traditional Chinese characters, which would be more commonly used in Hong Kong or Taiwan. Some extenuating combinations (*e.g.*, “官网注册 + 棋牌游戏”, “official website registration + card games”) or technical content (*e.g.*, “群发机”, “group messaging machine”) may violate QQMail's Terms of Service against fraud and spam. Given the sensitive nature of the extenuating combinations, many would seem better suited as censored terms themselves instead of instruments for disabling censorship. We discuss possible explanations for these perplexing findings in Section 6.

5 OTHER FINDINGS

In this section we investigate other aspects of QQMail's censorship system.

5.1 Are there exceptions to the exceptions?

One might imagine that, just as the presence of a censored keyword combination's extenuating combinations can prevent censorship, these extenuations may themselves have exceptions whose presence would again enable censorship. Alternatively, there may be some machine learning or natural language processing algorithm holistically evaluating the entire message for sensitivity.

To test this hypothesis, for five days, we generated samples by randomly choosing a censored keyword combination and one of its extenuating combinations and then completing the sample using the keyword-list-sourced method. If the sample was censored, we then isolated which new combination of keywords was responsible. However, we found no new censored keyword combinations that were not also censored when sent in isolation, suggesting there are no exceptions to the exceptions or any other more complicated system responsible for the censorship that we are measuring.

5.2 Does the number of recipients affect censorship?

For one day, we tested for censored keyword combinations sending to ten QQMail recipients instead of one. We found no new censored keyword combinations that were not also censored when sending to one recipient.

5.3 Does censorship vary with MX server?

Our tests ordinarily connect to mx3.qq.com, the lowest-numbered (thus highest) priority server in the MX records for qq.com. However, one might hypothesize that this MX server, located in Hong Kong, may implement less censorship than others, especially if it is intended for international email.

To test this hypothesis, we repeated the MX lookup, instead using 1.2.4.8, a DNS server in mainland China, receiving the same MX records as before. We then repeated the address lookup for mx3.qq.com using 1.2.4.8, this time receiving the IP address of a server in Guangdong, China. We tested using this MX server for one day. However, we found no censored keyword combinations that were not also censored when sent to the original Hong Kong-based MX server.

5.4 How were censored keyword combinations derived?

Although we cannot estimate the total number of censored keyword combinations, nothing about 173, the number of known censored keyword combinations, suggests that this list was not curated by hand. Moreover, among the 101 single-component censored combinations, the first and third longest, 就中国政府公布第三批恐怖嫌疑人名单谣言声明 (statement on rumors of the third batch of terrorist suspects released by the Chinese government) and 中共的侵略及维吾尔族的现状 (the invasion of the Chinese Communist Party and the current situation of the Uyghurs) do not appear in any other lists sourced in our keyword-list-sourced sampling. However, when searched using Google, the only results point to a keyword blacklist (https://raw.communitydragon.org/9.20/game/data/languagefilters/filter_zh_cn.txt) used by an online Website maintained by Tencent subsidiary Riot Games. While this suggests that some of QQMail's censored terms are censored across other Tencent products, we found that most of QQMail's censored combinations have little in common with any of the known keyword blacklists from Tencent products including WeChat.

5.5 Are our measurements representative?

To confirm that our measurements presented in this paper reflect the real-world experience of QQMail users, we performed the following tests between a Gmail and a QQMail test account, each using their respective Web UIs. From the QQMail account to the Gmail account, we sent (1) 20 messages each containing a randomly chosen censored keyword combination, (2) 20 messages each containing a randomly chosen censored keyword combination and one of its extenuating combinations, and (3) 20 messages each containing a random sample generated using the keyword-list-sourced method (see Section 3.2). We also sent (4) 20 messages containing a randomly chosen censored keyword combination from the Gmail account to the QQMail account. All messages in cases (1) and (4) were censored, and all messages in cases (2) and (3) were not censored, as expected.

In (2) and (3), we sent each email to a unique address by appending to the address's user name a plus sign and a unique string, a Gmail feature to create email aliases [9]. We did this as, incidental to this testing, we discovered another perplexing aspect to QQMail's automated censorship system, which is that if a QQMail user has previously sent an email to an address, then the automated censorship system would no longer censor email between that QQMail user and that other address. We are not currently sure how recently the user must have previously sent an email, or if there is any time limitation. Although such a feature may make sense for blocking spam, none of the censored keyword combinations we discovered were related to spam.

5.6 Side-channel vulnerability

Utilizing the censorship behavior described above, we developed a side-channel technique to measure whether a

QQMail user had previously sent an email to an external email address (our testing was with a Gmail address). Namely, by spoofing an email containing censored content from (*e.g.*) example@gmail.com to the QQMail user, we can determine whether the QQMail user has previously sent email to example@gmail.com by the MX server's response. If it is 550 "Mail content denied", then the QQMail user has not previously sent email to the external address. Conversely, if the response is 250 "OK", then the QQMail user has. As before, we are not currently sure how recently the QQMail user must have previously received an email, or if there is any time limitation, but we found that the period can be at least 24 hours.

We disclosed this vulnerability to Tencent, who responded: "[T]he QQMail anti-spam strategy will judge the degree of maliciousness based on multiple dimensions. Mail exchanges between users are one of the factors that reduce the degree of maliciousness, but not all. Therefore, it is not possible to accurately deduce whether there are email exchanges between users based on whether the email content is denied." As of July 2021, we can no longer reproduce this issue.

6 DISCUSSION

It is unclear why in our testing the censorship system fails to censor email from external addresses to whom the receiving QQMail user has previously sent email. This feature would make sense for blocking spam, where a QQMail user having previously sent email to an external address would signal that the user does not view that address as a spammer's. To support this hypothesis, there exists documentation that Voice of America's Chinese branch sends mass unsolicited email to Chinese users concerning censorship evasion tools [24], which could explain how political terms would end up blocked as spam. Moreover, China's National Internet Emergency Center warned that "reactionary information spammed by hostile forces poses great threats to political and social stability." [27] However, despite our sampling sources including references to phishing, pornography, and illicit goods, we did not find any of these topics censored. The censored keyword combinations we discovered were either politically motivated, including motivations such as blocking racially discriminating content against Uyghurs, or motivated by protecting the Tencent company.

More perplexing is the counterintuitively sensitive nature of QQMail's extenuating keyword combinations, for which we can provide a few possible explanations. First, this behavior could be a deliberate design to allow for sensitive information to be disseminated among users, which makes dissidents and potential mobilization efforts visible to the company. Qin *et al.* [18], who found "a shockingly large number of posts on highly sensitive topics" on Chinese social media, argue that this is so that the government can gauge and surveil public sentiment. We think that this explanation is unlikely as we found no pattern as to how each extenuating combination pairs with a censored keyword combination. Rather, past research on Chinese social media censorship suggests that companies tend to consider censorship requirements a mundane yet necessary task to operate a business without attracting official reprimands in China, where companies may simply copy and paste keyword lists from one another [13] or defy censorship for commercial reasons [17].

A second explanation is that this behavior could be a technical oversight by QQMail's developers. The findings in Section 5.6 suggest that other technical oversights may exist in their censorship system. For instance, if the extenuating keywords are part of a more traditional spam detection system and the censored keywords are part of a censorship system, then the presence of spam keywords may short-circuit the detection of censorship keywords. However, then it is unclear why censored keyword combinations have different extenuating keyword combinations.

A third explanation is that this behavior is an artifact of some greater system, which, if we understood it in its entirety, then these findings would seem intuitive. However, our findings in Section 5.1 failed to show that this system is any more sophisticated than a set of censored keyword combinations each having its own set of other combinations whose presence disables their corresponding censored keyword combination's censorship.

Regardless of the intention of these extenuating combinations, our work shows that one must be aware of their existence in order to correctly identify which combination of keywords is triggering the censorship of a message.

7 FUTURE WORK

While our work focuses on the censorship of text content in email, more work is needed to investigate MIME (multipurpose Internet mail extensions) censorship, including that of HTML content or that of image, audio, video, or document attachments. Further, while our work focuses on QQMail, another direction for future work is to explore using this measurement technique against other email providers in China, which, given previous work [13, 15, 16, 21], we would predict to have diverging implementations of censorship.

Finally, our work found that algorithms naive to the presence of extenuating combinations may silently produce censored keyword combinations with spurious features. Future researchers studying keyword censorship should investigate whether such exceptions occur in the system which they are studying and, if so, use an algorithm which is robust to them.

ACKNOWLEDGMENTS

We thank the anonymous reviewers for their helpful feedback. This project was supported by Open Society Foundations. This project was supervised by Ron Deibert, Citizen Lab Director.

AVAILABILITY

The censored keyword combinations that we discovered and each of their extenuating combinations are available here: https://github.com/citizenlab/chat-censorship/qqmail

REFERENCES

[1] Anonymous. 2014. Towards a Comprehensive Picture of the Great Firewall's DNS Censorship. *USENIX Workshop on Free and Open Communications on the Internet* (2014). https://www.usenix.org/system/files/conference/foci14/foci14-anonymous.pdf

[2] Miguel Araujo. 2017. What is QQ Mail and what is it for? Available at https://www.qqmail.info/what-is-qq-mail-and-what-is-it-for.html. (2017).

[3] Kevin Bock, George Hughey, Louis-Henri Merino, Tania Arya, Daniel Liscinsky, Regina Pogosian, and Dave Levin. 2020. Come as You Are: Helping Unmodified Clients Bypass Censorship with Server-side Evasion. In *Proceedings of the Annual conference of the ACM Special Interest Group on Data Communication on the applications, technologies, architectures, and protocols for computer communication*. 586–598.

[4] Carolyn Clancy. 2020. Country overview: China. Available at https://help.returnpath.com/hc/en-us/articles/220562707-Country-overview-China-. (2020).

[5] Jedidiah R. Crandall, Masashi Crete-Nishihata, Jeffrey Knockel, Sarah McKune, Adam Senft, Diana Tseng, and Greg Wiseman. 2013. Chat program censorship and surveillance in China: Tracking TOM-Skype and Sina UC. *First Monday* 18, 7 (6 2013). http://firstmonday.org/ojs/index.php/fm/article/view/4628/3727

[6] Jedidiah R. Crandall, Daniel Zinn, Michael Byrd, Earl Barr, and Ric East. 2007. ConceptDoppler: A weather tracker for Internet censorship. In *14th ACM Conference on Computer and Communications Security, Oct.29-Nov2, 2007*. 1–18.

[7] James Doubek. 2018. China Removes Presidential Term Limits, Enabling Xi Jinping To Rule Indefinitely. (3 2018).

[8] Barney Glaser and Anselm Strauss. 2006. *The Discovery of Grounded Theory: Strategies for Qualitative Research*. INternational and Pan-American Copyright Conventions, New Jersey.

[9] Google. [n. d.]. Create task-specific email addresses. Available at https://support.google.com/a/users/answer/9308648?hl=en. ([n. d.]).

[10] Seth Hardy. 2013. *Asia Chats: Investigating Regionally-based Keyword Censorship in LINE*. Technical Report. Citizen Lab, University of Toronto. https://citizenlab.ca/2013/11/asia-chats-investigating-regionally-based-keyword-censorship-line/

[11] Nguyen Phong Hoang, Arian Akhavan Niaki, Jakub Dalek, Jeffrey Knockel, Pellaeon Lin, Bill Marczak, Masashi Crete-Nishihata, Phillipa Gill, and Michalis Polychronakis. 2021. How Great is the Great Firewall? Measuring China's DNS Censorship. (2021). arXiv:cs.CR/2106.02167

[12] Gary King, Jennifer Pan, and Margaret Roberts. 2013. How Censorship in China Allows Government Criticism but Silences Collective Expression. *American Political Science Review* 107, 2 (2013), 326–343.

[13] Jeffrey Knockel, Masashi Crete-Nishihata, Jason Q. Ng, Adam Senft, and Jedidiah R. Crandall. 2015. Every Rose Has Its Thorn: Censorship and Surveillance on Social Video Platforms in China. In *5th USENIX Workshop on Free and Open Communications on the Internet*.

[14] Jeffrey Knockel, Masashi Crete-Nishihata, and Lotus Ruan. 2018. The effect of information controls on developers in China: An analysis of censorship in Chinese open source projects. In *First Workshop on NLP for Internet Freedom*. https://www.aclweb.org/anthology/W18-4201

[15] Jeffrey Knockel, Lotus Ruan, and Masashi Crete-Nishihata. 2017. Measuring Decentralization of Chinese Keyword Censorship via Mobile Games. In *7th USENIX Workshop on Free and Open Communications on the Internet*. https://www.usenix.org/conference/foci17/workshop-program/presentation/knockel

[16] Rebecca MacKinnon. 2009. China's Censorship 2.0: How companies censor bloggers. *First Monday* 14, 2 (2009). http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/2378/2089

[17] Blake Miller. 2019. The Limits of Commercialized Censorship in China. (2019). https://doi.org/10.31235/osf.io/wn7pr

[18] Bei Qin, David Strömberg, and Yanhui Wu. 2017. Why Does China Allow Freer Social Media? Protests versus Surveillance and Propaganda. *Journal of Economic Perspectives* 31, 1 (2017), 117–140.

[19] R Rambert, Z Weinberg, D Barradas, and N Christin. 2021. Chinese wall or Swiss cheese? Keyword filtering in the Great Firewall of China. In *Proceedings of the 30th Web Conference (WWW'21)*. Ljubljana, Slovenia (online).

[20] Lotus Ruan, Jeffrey Knockel, Jason Q. Ng, and Masashi Crete-Nishihata. 2016. *One App, Two Systems: How WeChat uses one censorship policy in China and another internationally*. Technical Report. Citizen Lab, University of Toronto. https://citizenlab.ca/2016/11/wechat-china-censorship-one-app-two-systems/

[21] Taiyi Sun and Quansheng Zhao. 2021. Delegated Censorship: The Dynamic, Layered, and Multistage Information Control Regime in China. *Politics \& Society* (2021), 00323292211013181.

[22] Tokachu. 2006. The Not-So-Great Firewall of China. *2600* 23, 4 (2006), 58–60.

[23] United Nations. 2011. *EASTERN TURKISTAN ISLAMIC MOVEMENT*. Technical Report.

[24] United States Department of State and the Broadcasting Board of Governors Office of Inspector General. 2010. Report of Inspection: Voice of America's Chinese Branch. Available at https://www.stateoig.gov/system/files/145823.pdf. (2010).

[25] Ruohan Xiong and Jeffrey Knockel. 2019. An Efficient Method to Determine which Combination of Keywords Triggered Automatic Filtering of a Message. In *9th USENIX Workshop on Free and Open Communications on the Internet*. Santa Clara.

[26] Tao Zhu, David Phipps, Adam Pridgen, Jedidiah R Crandall, and Dan S Wallach. 2013. The Velocity of Censorship: High-fidelity Detection of Microblog Post Deletions. In *Proceedings of the 22nd USENIX Conference on Security*. 227–240.

[27] 国家互联网应急中心. 2003. 中国互联网反垃圾邮件的现状. (9 2003). https://www.cert.org.cn/publish/main/49/2012/20120330182937909855444/20120330182937909855444_.html

[28] 秦勉. 2015. 详述GFW对SMTP协议的三种封锁手法. Available at https://fqrouter.tumblr.com/post/43400982633/gfw smtp . (2015).

Appendices are supporting material that has not been peer-reviewed.

A SUPPLEMENTARY ALGORITHMS

In this section we include an algorithm referenced in Section 3.3.

Algorithm 1 CABSE(C_0, s)

```
$C \leftarrow \{\}$
repeat
    $i \leftarrow$ BinSearchE($C_0 \cup C, s$)
    $j \leftarrow i + 1$
    $k \leftarrow |s|$
    while $j < k$ do
        if IsCensored($C_0 \cup C \cup \{s[i{:}j], s[i+1{:}]\}$) then
            $j \leftarrow j + 1$
        else
            $k \leftarrow j$
        end if
    end while
    $C \leftarrow C \cup \{s[i{:}j]\}$
    if $j \neq |s|$ then
        $s \leftarrow s[i+1{:}]$
    else
        $s \leftarrow$ ""
    end if
until $|s| = 0$ or not IsCensored($C_0 \cup C$)
return $C$
```

Algorithm 2 BinSearchE(S, g)

```
$lo \leftarrow 0,\ hi \leftarrow |g|$
while $hi - lo > 1$ do
    $mid \leftarrow \lfloor (lo + hi)/2 \rfloor$
    if IsCensored($S \cup \{g[mid{:}]\}$) then
        $hi \leftarrow mid$
    else
        $lo \leftarrow mid$
    end if
end while
return $lo$
```

As described in Section 3.3, *CABSE*, an algorithm to isolate *extenuating* keyword combinations, is a modified version of *CABS* as described in Xiong *et al.* [25] which was for isolating *censored* keyword combinations. The argument C_0 is a set of keyword components of some censored keyword combination, and s is a string which, when present in a message with C_0, deactivates C_0's censorship.