Every Rose Has Its Thorn: Censorship and Surveillance on Social Video Platforms in China

Jeffrey Knockel[1,2], Masashi Crete-Nishihata[2], Jason Q. Ng[2], Adam Senft[2], and Jedidiah R. Crandall[1]

[1]*Dept. of Computer Science, University of New Mexico*
[2]*Citizen Lab, Munk School of Global Affairs, University of Toronto*

Abstract

Social media companies operating in China face a complex array of regulations and are liable for content posted to their platforms. Through reverse engineering we provide a view into how keyword censorship operates on four popular social video platforms in China: YY, 9158, Sina Show, and GuaGua. We also find keyword surveillance capabilities on YY. Our findings show inconsistencies in the implementation of censorship and the keyword lists used to trigger censorship events between the platforms we analyzed. We reveal a range of targeted content including criticism of the government and collective action. These results develop a deeper understanding of Chinese social media via comparative analysis across platforms, and provide evidence that there is no monolithic set of rules that govern how information controls are implemented in China.

1 Introduction

The Chinese Internet is a complex ecosystem that includes multiple layers of technical and regulatory information controls that are affected by actors with different positions of influence and responsibility. Extensive work has been done on China's national-level Internet filtering system, but this system is only one of many layers of information control in the country.

Gaining a wider understanding of censorship and surveillance in China requires analysis of its Internet platform developers and companies. These companies operate in a highly constrained regulatory environment in which they are responsible for the content on their services and subject to fines and loss of operating licenses if they are found in violation. This governance model effectively pushes responsibility for information control to the private sector.

Previous work has shown inconsistencies in how different Chinese Internet companies conduct censorship, which suggests there may be general directives provided by the government on what content to censor, but the companies have a degree of flexibility for how they implement it [16, 25, 31]. Further exploring these observations is challenging due to methodological constraints. The majority of studies on Chinese social media rely on sample testing in which researchers develop a set of content suspected to be blocked by a platform, send the sample to the platform, and record the results. This approach introduces inherent bias as results are only as accurate as the overlap between the sample and the actual content filtered. Another approach is to observe changes to a system (such as content deletion) over time. Studies using this method are typically limited to snapshots within a specific period, which constrains longitudinal analysis. Applications that implement censorship and surveillance on the client-side (*i.e.*, by the application itself rather than on a remote server) present a unique research opportunity. Reverse engineering applications can reveal keyword lists used to trigger censorship and surveillance. These lists are unbiased samples that provide comprehensive visibility into technical implementation and target content.

This study provides a broad look into keyword surveillance and censorship across social video platforms (SVPs), a popular class of applications in China. SVPs combine real-time video streaming and social networking features that enable users to broadcast content and create interactive groups. One of the most popular uses is broadcasting karaoke performances. SVPs are primarily monetized through the sale of virtual goods (such as virtual roses) that users give to performers during broadcasts. While musical performances account for the majority of revenues, SVPs are expanding to gaming, education, financial analysis, and online dating applications.

Through reverse engineering, we identify client-side keyword censorship in four of the most popular SVPs: YY, 9158, Sina Show, and GuaGua. In the case of YY we also find keyword surveillance capabilities. Our analysis reveals a dataset of 17,547 unique keywords, which

Company	Product	Reg. Users	MAUs
YY Inc.	YY	861.4 mn.	117.4 mn.
Tian Ge	9158	245.0 mn.	14.4 mn.
	Sina Show		
Jinhua Changfeng	GuaGua	70 mn.	not available

Table 1: **Social Video Users by Platform**

to our knowledge is the largest unbiased collection of censorship keywords currently available. [1] Our main findings are as follows:

Inconsistencies in targeted content and implementation between platforms: Comparing our SVP dataset to previously collected chat client censorship keyword lists [16] allows for the first comparison of unbiased keyword samples across different industry segments. Our analysis reveals limited list overlap between companies, which substantiates previous findings that suggest companies are only given general directives from authorities and have a degree of flexibility in the implementation.

Range of targeted content including criticism of the government and collective action: While there is limited direct overlap in unique keywords, across lists we see trends in the topics that are targeted including social issues, criticism of the government, and collective action. These findings serve as a counterpoint to previous work from King *et al.* [21, 22] who posit that content related to collective action is heavily censored on Chinese social media while content critical of the government is often allowed to persist.

Our findings provide strong evidence that there is no monolithic set of rules governing how information controls are implemented in China and that developing holistic understandings of the Chinese Internet requires comparative analysis across platforms.

2 Background

The most popular SVPs in China include YY, Sina Show, 9158, and GuaGua. YY is developed by YY Inc. based in Guangzhou, China, and is the largest platform in terms of user population. As of December 2014, YY had 861.4 million registered users and 117.4 million average monthly active users (MAUs) [6]. In November 2012, YY Inc. announced an initial public offering on the Nasdaq stock market. It is currently the only Chinese SVP company to be traded on the US stock market.

Tian Ge Interactive Holdings Limited based in Hangzhou, China owns and operates two SVPs: 9158 and Sina Show. In 2010, Sina Corporation invested 10 million dollars (representing a 25% stake) in Tian Ge and provided the company a sole license for the operation of of Sina Show. Tian Ge reports user numbers as aggregates across its platforms and in 2014 had 245 million registered uses and 14.4 million MAUs [3, 4]. In July 2014, Tian Ge went public on the Hong Kong Stock Exchange.

Jinhua Changfeng Information Technology Co., Ltd., is a privately held company based in Zhejiang Province, China that provides the GuaGua platform, which as of 2013 had 70 million registered users [9].

See Table 1 for a breakdown of user bases across SVPs.

2.1 Legal and Regulatory Environment in China

The Communist Party of China (CPC) attempts to balance the growth of information and communication technologies in the country and limits on speech that can threaten its power [26]. In 2010 China's State Council Information Office published what is considered the first government issued policy document on the Internet. It includes a list of prohibited topics:

> endangering state security, divulging state secrets, subverting state power and jeopardizing national unification; damaging state honor and interests; instigating ethnic hatred or discrimination and jeopardizing ethnic unity; jeopardizing state religious policy, propagating heretical or superstitious ideas; spreading rumors, disrupting social order and stability; disseminating obscenity, pornography, gambling, violence, brutality and terror or abetting crime; humiliating or slandering others, trespassing on the lawful rights and interests of others; and other contents forbidden by laws and administrative regulations [8].

Companies are held liable for content on their platforms and are expected to invest in staff and technology for ensuring compliance with government regulations. Failure to comply with regulations can lead to fines or revocation of operating licenses. Complicating matters is the vague language used in Chinese regulatory documents. Terms like "disrupting social order and stability" are not clearly defined and punishments are meted out seemingly arbitrarily, thus pushing companies and users to both over-censor and self-censor—a phenomenon coined by Perry Link as "the anaconda in the chandelier" [24].

In public filings for YY and Tian Ge both companies underline the risk that their businesses face from potential legal sanctions being brought against them for hosting prohibited content [2, 6]. The companies also highlight the risk of being affected by government campaigns such as "Clean the Web 2014," which was an government effort to crack down on the creation and dissemination of

pornographic content online. During this campaign, Sina Corporation received notices regarding prohibited content on its platforms and was subsequently fined 5.1 million RMB, had licenses temporarily revoked, and saw its stock price drop as a result. This campaign demonstrates the dynamic nature of Internet regulations in China, and shows companies are subject to unpredictable enforcement, which can impact their bottom line.

Unlike most other social media platforms in China, SVPs have an added dimension of sharing revenues with performers. This business model and the general SVP user experience encourages performers to keep audiences engaged and spending on virtual goods. The popularity of SVPs, the diversity of real-time media content, and the virtual goods business model puts these platforms under particular pressure to monitor and manage user activity.

2.2 Content Monitoring and Censorship on Social Video Platforms

To comply with China's laws and regulations SVPs manage content through a combination of terms of service (TOS), automated content monitoring and filtering systems, and dedicated review teams.

YY has an extensive TOS that includes descriptions of prohibited content and a five-level system for penalties that range from freezing the account from 7 days (level 1), 30 days (level 2), 120 days (level 3), 360 days (level 4), or permanently (level 5). Serious violations that warrant a level 4 or 5 response include publishing pornography; publishing content that endangers national security or undermines national unity, social stability, or national religious policy. Offences that carry lower level punishments include vulgar jokes, verbally abusing other users, and copyright infringement [6]. Performers are expected to obey an additional list of regulations that include the same high level prohibitions and other specific guidelines on inappropriate attire and performance material. Failure to comply can result in fines and account suspensions [7].

To enforce these TOS, YY has a team within the data security department that maintains "24-hour surveillance" on content and is supported by a system that periodically "sweeps" the platform for offensive content and "automatically" filters keywords. The company also describes a voice monitor system that provides "various alerts on sensitive words or abnormal activities of users, channels, or groups" [6].

Tian Ge has a similar combination of controls. It employs a team of 74 content monitors who identify TOS violations and enforce internal policies. In public filings the company describes an image processing system used to detect skin tone and facial features to flag nudity or "sexually suggestive partial nudity." Screenshots of video chat rooms are randomly captured every 1-3 minutes and processed through the detection system. Flagged content is then sent to the content monitoring team for further review. The company also generally describes audio monitoring and keyword filtering systems. In addition, it provides the same level of access that its content monitoring team has to the Jinhua City Municipal Public Security Bureau to allow the authorities means to "monitor and supervise the activities" on the platform [2].

As GuaGua is a private company less information is available on its internal operations. In a 2013 interview, co-founder Dong Guanjie claims the platform has a content management team of over 100 staff [9]. GuaGua's TOS explains the company performs automated and manual inspection of content and deletes any infringements [1]. Similar to YY and Tian Ge emphasis is placed on incentives for user self-regulation and financial penalties for violations.

3 Related Work

The majority of research on Internet censorship in China focuses on the *Great Firewall of China*, its national-level Internet filtering system [12, 15, 17, 19, 29, 34, 35, 37]. In comparison to this literature, the number of studies on surveillance and censorship on Chinese social media is limited.

The microblogging service, Sina Weibo, has been the focus of a number of studies that show the dynamic nature of content filtering on the platform. Bamman, *et al.* [13] conducted statistical analysis of deleted Weibo posts and found that posts with sensitive words and from certain geographic locations (*e.g.*, Tibet and Qinghai) have a higher deletion rate. Zhu, *et al.* [36] measured censorship on Weibo and found that retroactive post deletions occur within minutes and the censors use a variety of automated tools. The University of Hong Kong has developed WeiboScope, a data collection and visualizations system for tracking censorship on Weibo [5]. Fu *et al.* [33] use this system to show that real name registration policies on Weibo may have caused some users to self-censor. Ng [27] identified numerous ways that a Weibo message could be censored or held in review both before and after being posted, confirming the usage of both automated and manual review processes.

King, *et al.* [21] collected posts from 1,382 Chinese social media Web sites and, through statistical analysis comparing censored and uncensored posts, contend that censorship focused on content that represented, reinforced, or encouraged collective action.

Previous work has found inconsistencies in censored content and the technical implementation of censorship across services. MacKinnon [25] examined 15 different Chinese blog providers and found significant variation in the extent and implementation of content censorship. Villeneuve [31] analyzed keyword filtering in search engines

localized for the Chinese market and found a similar lack of overlap in censored keywords. These studies suggest that content filtering across platforms is highly decentralized and affords companies a level of flexibility in implementing controls.

The previously outlined studies relied on testing samples [25, 27, 31] or observing changes [13, 21, 36] (*e.g.* deletions) in a subset of content over a fixed period. Other work focused on client-side implementations of censorship and surveillance have extracted unbiased keyword lists used to trigger these functions and analyzed changes over time. Villeneuve [32] revealed keyword surveillance in TOM-Skype by discovering chat logs uploaded by the client through an insecure publicly accessible Web server hosted in China. Knockel, *et al.* [23] reverse engineered multiple versions of TOM-Skype, decrypted censorship and surveillance keyword lists and reported on updates over a one month period. In Crandall *et al.* [16], censorship keyword lists were obtained from Sina UC (another chat client used in China) and compared to the TOM-Skype lists, tracking each for a period of over 22 months, categorizing the keywords into granular content categories, and correlating list updates with current events. Similar to [25, 31] this study revealed inconsistencies in the implementation of keyword censorship between the programs and found only 3% overlap in unique keywords within the total dataset of 4,256 keywords.

Hardy [20] reversed engineered LINE, a mobile chat client developed by a Japanese company and marketed to countries around the world including China, and revealed regionally-based keyword filtering implemented on the client-side that is enabled for users with accounts registered to mainland China phone numbers. Analysis of these keyword lists revealed limited overlap with TOM-Skype and Sina UC [18].

4 Technical Analysis

In this section, we describe the technical implementations of keyword censorship and, if present, keyword surveillance in each of the SVPs we analyzed.

We found that three Chinese SVPs not described in this paper, VV (*51vv.com*), Sixroom (*6.cn*), and BoBo (*bobo.com*), perform keyword censorship on the server-side. We determine that an application censors on the server-side by first ensuring that there are no obvious signs of a client-side censorship implementation such as one of its censored keywords appearing in plain text in any of the application's files. Then we compare the packet trace of sending a censored keyword (*e.g.*, "falun") with sending a nearly identical uncensored keyword ("galun") and, if the application censors by asterisking out sensitive keywords, with sending that keyword asterisked out ("*****"). If the first comparison is repeatedly similar and, if performed, the second comparison is less so, we conclude that the censorship is server-side. If the application displays a warning message upon entering a censored message, we also ensure that the warning message does not display when the application has no access to the Internet. We leave it as future work to analyze the server-side censorship implementations used by these platforms.

4.1 YY Censorship and Surveillance

YY 7.1 downloads three different keywords lists with the following names: *Finance*, *Normal*, and *High*.

The *Finance* keyword list is downloaded from http://do.yy.duowan.com/financekwordlist. These keywords are downloaded in plain text in UTF8-encoded XML. Keywords in this list are related to phishing scams. When a user receives a keyword from the list the following warning message is displayed in the chat window: "YY安全提示：聊天中若有涉及财产的操作，请一定要先核实好友身份，谨防受骗!" (YY Security Tip: This chat seems to involve managing assets; please be sure to verify the identity of a friend to avoid being cheated!).

The *Normal* keyword list is downloaded from http://do.yy.duowan.com/NormalKWordlist.txt as a base64-encoded list of UTF16-encoded keywords each separated by a carriage return followed by a line feed. If an outgoing or incoming message contains a keyword from this list, those keywords are asterisked out in the chat window. In the case of an outgoing message, those keywords are also asterisked out in the message sent over the network.

The *High* keyword list is downloaded from http://do.yy.duowan.com/HighKWordlist.txt. Like the *Normal* list, it is a base64-encoded list of UTF16-encoded keywords each separated by a carriage return followed by a line feed. If an outgoing message contains a keyword from this list, that message is silently filtered. If an incoming message contains a keyword from this list, it appears in the chat window as a blank message.

4.1.1 YY Surveillance

The keywords from both the *Normal* and *High* lists are also used to trigger surveillance. When attempting to send a message containing keywords from either of these lists, a surveillance message is sent via an HTTP GET request to a URL of the form:

http://sere.hiido.com/do.action?
id=**<id>**&content=**<content>**

<id> is a hex encoding of a hash computed as

md5(⌊**<seconds since unix epoch>**/1000⌋+
`";username=report"+`
`";password=pswd@1234")`.

Note that the username and password in the hashed string are hardcoded; these are not the username and password

of the sender or receiver of the triggering message. **<content>** is a base64 encoding of the following string:

type=2;uid=**<sending user id #>**;touid=**<receiving user id #>**;keyword=**<triggering keyword>**;txt=**<entire triggering message>**

Type is hardcoded to 2.

4.2 Sina Show Censorship

Sina Show 3.4 comes installed with keyword lists for censoring messages and also downloads additional keywords remotely from its servers. When an outgoing message is censored, the message is not sent, and the following warning message is displayed in the chat window: "系统过滤，你发送的信息含有非法字符，请重新输入！" (System filter: the message you sent contains illegal words; please re-enter!) When an incoming message is censored, the contents of the incoming message are replaced by the following message in the chat window: "发送的消息有非法词汇，已经被自动屏蔽"(The message sent contains illegal vocabulary; it has been automatically blocked.)

Sina Show comes installed with a binary database of keywords in a file named `Word_410.ucw` and downloads updates for it from http://www.51uc.com/uc_interface/down_policy/Word_410.ucw. This file is a custom binary container storing sensitive GBK-encoded keywords that have been encrypted using Blowfish in ECB mode with the 8-byte key `Dey,1blE`. A standard library implementation of Blowfish cannot be used to decrypt these keywords, however, as the Blowfish implementation used by Sina Show is atypical, containing byte endianness inconsistencies and bit shift discrepancies compared to the standard implementation.

Each keyword in this file is also associated with a category number from 1 to 8, inclusive, or 12. For this reason, keywords often appear more than once in this file if they belong to multiple categories. However, Sina Show at present only utilizes category 5, which it uses to censor chat messages. If the original purpose of the additional categories was like that found in Sina UC [16], it may be the case that the other categories were originally used to censor usernames or other strings in an older version of the program. As of May 11, 2015, 888 of the 2709 keywords in this file are in category 5.

Sina Show also has GBK-encoded lists of keywords included in plain text built into many of its binaries. `SinaShow.exe` contains a list of 1224 keywords, which are also included in `ChatRoom.dll` and `Props.dll`. `UCClient.dll` also includes another list of 910 keywords. However, among all of these built-in keywords, only 108 keywords, the 1114th through the 1221st keywords in `SinaShow.exe` are actually referenced by any binary code, and these are also used to censor chat messages. The other unused keywords may correspond to presently unused categories as we saw with the downloaded keywords.

4.3 9158 Censorship

9158 6.9 is installed with two lists of keywords, `filnick.xml` and `filter.xml`. Although these XML files self-identify as being GB2312-encoded, they are really GB18030-encoded. The former list is used to replace sensitive keywords with asterisks in user names, whereas the latter list is used to replace sensitive keywords with asterisks in both outgoing and incoming chat messages. Updates to the latter list are also downloaded from http://mimtenroom.9158.com/web9158/filter.zip. The `filter.xml` file also includes a version number of the list, which is an integer in the hundreds that we have found to increase by a few every time the list is updated. The version number of the list installed with the program, however, is greater than the version number in any of the updates we have seen offered for download, suggesting that the sequence may have reset or forked at some point.

In addition to keyword censorship, we found that if a chat message contains six or more English alphabet letters, then all of its English alphabet letters are asterisked out. The intent of this filtering is not clear. Aside from stifling English conversation, this may be intended to filter out URLs. Given that their keyword lists filter keywords like *http*,*www*, and *com*, it would seem they intend to filter all URLs.

4.4 GuaGua Censorship

GuaGua 6.2.38 has keywords built into `RuleCenterPlug.dll`. These keywords appear in plain text, GBK-encoded. Any outgoing message containing one of these keywords is never sent and the following warning message is displayed in the chat window: "消息发送失败,含有违法或不文明字符!" (Failed to send message; it may contain illegal or uncivilized words!) GuaGua does not filter incoming messages.

5 Keyword Analysis

In total, our dataset consists of 42 lists, which together contain 17,547 unique keywords. The lists range in size from 20 to 13,244 unique keywords.

Table 2 shows the size of each list in terms of total keyword count and the number of unique keywords. The YY *High* list is the largest list in our dataset, whereas, if we exclude YY's smaller lists, GuaGua's list is the smallest.

List	Keywords	Unique
YY *Finance*	48	18
YY *Normal*	20	20
YY *High*	13,482	13,242
9158 *Nick*	65	59
9158 *Chat*	318	318
Sina Show `SinaShow.exe`	1,224	910
Sina Show `UCClient.dll`	910	910
Sina Show *Downloaded*	3,711	3,206
GuaGua	58	58

Table 2: **List size (May 17, 2015)**

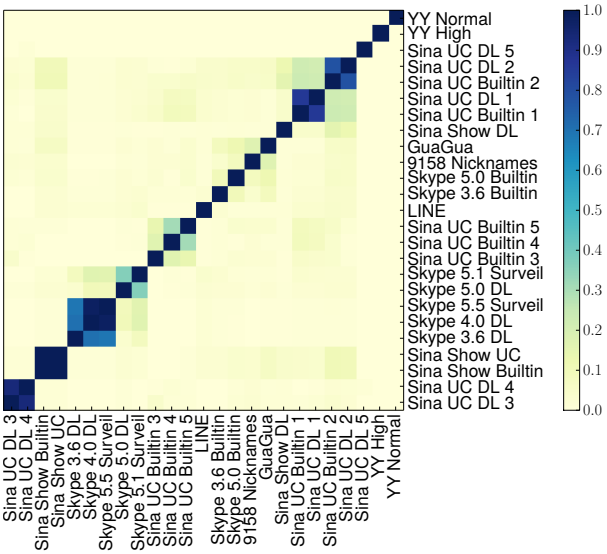

Figure 1: **Keyword lists clustered by Jaccard similarity**

5.1 Similarity Comparison Across Keyword Lists

In Figure 1, using the Nearest Point Algorithm, we hierarchically cluster SVP keyword lists along with the TOM-Skype and Sina UC lists [16] and the latest list from LINE [18] by Jaccard similarity coefficient (*i.e.*, the size of the intersection of two sets divided by the size of their union). Using this method, we find very little similarity between lists, and when lists are similar they are lists within the same company.

In Figure 2, we cluster the same keyword lists using a different similarity metric. We compute list x's similarity to y as $\max(\%$ of x in y, $\%$ of y in $x)$. The intuition behind this metric is that it would tease out lists that inherit from other lists. Although using this method we see more lists similar to each other within companies, lists from different companies remain mostly dissimilar with one exception: GuaGua is similar to many Sina Show lists. Closer inspection reveals that the GuaGua list is a near exact duplicate of a 2004-era list built into Sina UC that Sina Show's built-in lists build upon. The only difference in the GuaGua list is the addition of a single keyword. Both of the founders of GuaGua formerly worked on audio chat software at Langma UC (acquired by Sina Corporation in 2004 to become Sina UC) and Sina [9]. This employment history may explain why the GuaGua and Sina lists are so similar.

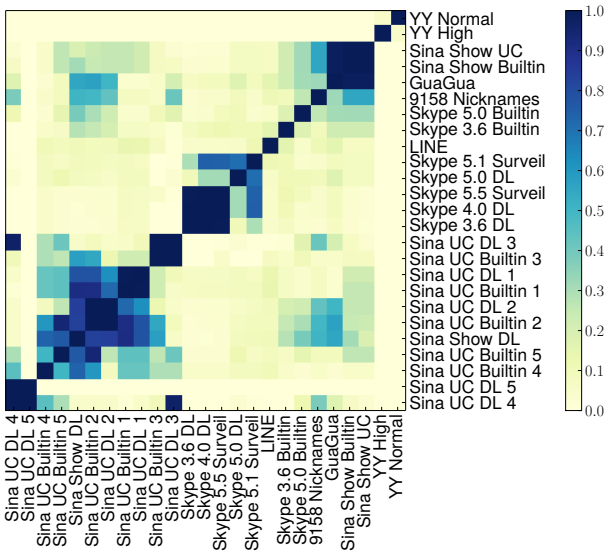

Figure 2: **Keyword lists clustered by** $\text{similarity}(x, y) = \max(\%$ of x in y, $\%$ of y in $x)$

5.2 Keyword Content Analysis

We used a combination of machine and human translation to translate the keywords to English and analyzed the context behind each one. Based on these translations and contextual information three researchers coded each keyword into one of 80 content categories grouped under six general themes based on a code book developed in [16]. We performed interrater reliability checks throughout the categorization process.

The six themes with example categories are: Social (prurient interests, illicit goods, gambling), Political (Communist Party of China, religious movements, ethnic groups), People (government officials, dissidents), Events (scheduled events, recurring events, current events), Technology (general technical terms, URLs, references to applications), and Miscellaneous (terms without clear context).

Our content analysis is on the 17,143 out of 17,547 keywords currently used to trigger censorship or surveillance events. This subset includes 7,371 URLs and URL fragments (such as *www*, *http*, *.B32.c*). These URLs cover a range of websites including phishing pages, pornography, independent media, and competing SVPs.

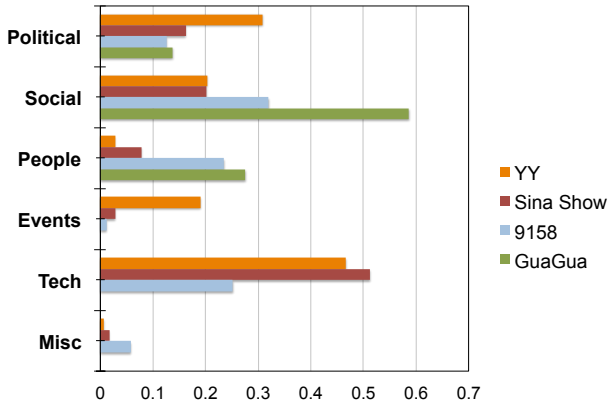

Figure 3: **Breakdown of list source by theme**

In the content analysis that follows we separate out URLs and report on 9,772 keywords.

Figure 3 shows a breakdown of keywords by theme across the four platforms (normalized by total number of keywords in each SVP). Keywords related to the social, political, and people themes are the most common across all four platforms, which shows that despite the lack of overlap in unique keywords between the platforms, they appear to be generally concerned about similar topics.

Social: The Social theme is divided into three categories: gambling and lottery (*e.g.*, online casinos, lottery games), illicit goods and services (*e.g.*, narcotics, firearms, counterfeit products), and prurient interests (*e.g.*, sexual interests, pornography, prostitution). All four SVPs included content related to prurient interests, whereas only 9158, Sina Show, and YY had content from the other two categories.

Political: The Political theme includes the widest range of content with 37 categories related to issues including the CPC, religious movements, ethnic minorities, and terrorism. Content related to the CPC covers both general references to the party and criticism of it. The only category found on all four platforms was Falun Gong (also the only political category on GuaGua's list). Notably 9158, Sina Show, and YY include keywords related to Uyghur issues, which account for the largest percentage of keywords within the political theme for Sina Show (45%) and YY (25%). YY *High* lists contain 0.5% Uyghur keywords in Arabic script that include references to terrorism and Islam. For example:

پارتلىقۇچ ياساش دەرسلىكى

which is a Uyghur phrase that translates to "instructions to create explosives."

Previously collected keyword lists do not include the Uyghur language or Arabic script [16, 18, 20]. The increased focus on Uyghur related content relative to previously available keyword lists may have been motivated by a June 2014 government campaign to censor terrorist content following attacks in the Xinjiang region. Thirty Chinese Internet companies signed a "letter of commitment" to block such content [11].

People: The People theme includes five categories: CPC officials, relatives and associates of CPC officials, dissidents, victims of crime, and names without clear context or identities. All four SVPs included reference to specific CPC officials including past leaders, "毛泽东" (Máo Zédōng), and current leaders, "习近平" (Xí Jìnpíng). Keywords often include varieties of homonyms for referring to leaders (*e.g.*, "习尽平", xí jǐn píng) and nick names such as Steamed Bun Xi ("习包子"), which refers to a photo that circulated online of Xi Jingping ordering lunch at a steamed bun shop that was subsequently criticized as a political show [10].

Events: The Event theme includes 23 specific events, 14 of which are related to protests and political mobilizations. Other event types include political forums (*e.g.*, CPC National Party Congress), and rumors around incidents (*e.g.*, the disappearance of Malaysia Airlines Flight 370). GuaGua had no references to events and 9158 only referenced the Bo Xilai scandal and the 2009 Urumqi riots. Sina Show included references to 11 events. YY referenced 18 events and had the greatest number of event related keywords (2,535). Over 90% of YY's event related keywords were references to the June 4 1989 Tiananmen Square Massacre (this category accounted for 32% of YY lists overall). We identify 8 events that are not present in the TOM-Skype / Sina UC dataset [16] including recent events such as the 2014 Occupy Central protests in Hong Kong, the 2014 Asia-Pacific Economic Cooperation Forum, and the 2014 terrorist attack in the Chinese city of Kunming. The addition of these recent events suggests sensitive events continue to be catalysts for censorship as seen in [16].

Technology: The technology theme has nine categories including a range of identifiers (phone numbers, QQ numbers and emails), which we suspect are primarily related to scams and illicit services. In other cases identifiers are clearly related to political issues (*e.g.*, emails related to Falun Gong media websites). Other categories include generic technical terms like Website ("网址"), and references to software used in China. Interestingly, three of the SVP keyword lists (YY, Sina Show, 9158) include references to other competing SVPs, which may be attempts to prevent users from being lured away from the provider's platform.

5.3 Keyword List Updates

Hourly data collection for the clients began on the following days: YY on February 7, 2015; 9158 on February 24, 2015; and Sina Show on March 11, 2015. (GuaGua does not download updates to its keyword list). We collected data through May 17, 2015. During this time, we saw 3

updates to YY *Normal*, 21 updates to YY *High*, and 8 updates to 9158 *Chat*; however, we saw no updates to Sina Show *Downloaded*, although its HTTP last-modified date would suggest that it was last updated February 9, 2015.

Two major updates to YY *Normal* occurred when names of Christian Chinese songs were added on April 23 and subsequently removed on April 30. The effort to block this content may have been in response to controversy over recent church demolitions [30]. Although religious songs are not explicitly prohibited by the YY TOS, it prohibits users from violating "Chinese religious policy" [7].

Updates to YY *High* were primarily in response to online scams. One update on February 9, however, was the addition of "bobo.com", a link to a competing SVP.

Updates to 9158 *Chat* included URLs (*dropbox.com*, *mediafire.com*) and keywords related to Islamic terrorism, and names of government officials. One official "周永康" (Zhou Yongkang) was added on May 6. On April 3, Zhou was charged with multiple corruption crimes, which led to a high profile trial in China [14].

6 Discussion and Conclusion

In this section we discuss the implications of our findings and how they may inform future work.

Inconsistencies in the content and implementation in keyword lists across companies and platforms: Previous studies that found inconsistencies in client-side keyword filtering (and in some instances surveillance) [16, 20] were limited to a small number of applications that were not very representative in terms of usage numbers. We substantiate these findings by analyzing exhaustive keyword lists across four of the top applications in an industry segment. We found very little consistency in the keywords censored by different applications (with the exception of GuaGua, whose founders had former employment with Sina). This finding is consistent with theories [24, 26] that companies are under vague pressure from the government to perform censorship and surveillance, leaving companies to decide how to implement it.

Targeted content relates to government criticism and collective action: King *et al.* [21, 22] acknowledge that social media sites in China have a flexible set of technical options for implementing censorship. However, they conclude that these diverse tactics lead to a mostly uniform outcome—content related to collective action (both pro- and anti-government) are heavily censored while "vitriolic blog posts about even the top Chinese leaders" are often allowed to persist in social media. This theory does not appear to align with the general regulations pushed to companies (*e.g.*, the prohibited topics listed in [8]), relies on sampling methods that may not comprehensively capture censored content, and is solely focused on take down of already published content in blog sites and bulletin board systems. Moreover, the study reports results as aggregates across the 1,382 sites that were analyzed, which may suggest greater uniformity than exists in practice. By contrast reverse engineering client-side implementations provides an exhaustive view into the exact content being targeted and the technical mechanisms that are utilized.

Our keyword content analysis brings the generalizability of King *et al.*'s theoretical arguments into question. Across the SVPs we studied and previous work on chat clients [16] we see keywords related to general discussion and criticism of the government and collective action. For example, all four SVPs include references to past and present CPC members. Apart from "Falun Gong" and "prurient interests" this is the only category shared by all SVPs. We observe criticism of the CPC that can be interpreted as collective action such as the Tuidang movement, which urges withdrawal from the party. Keyword lists also include general references to the CPC (*e.g.*, "共产党", Communist Party) and derogatory phrases (*e.g.*, "共匪", Communist gangsters). Inclusion of such keywords would limit both general and critical discussion of the government.

In the event theme we observe a high number of references to incidents related to collective action, but also those without a clear link. For example, references to the Asia-Pacific Economic Cooperation Forum include open-ended keywords (*e.g.*, "APEC") and specific criticisms such as "一开盛会就无霾" (once the summit starts, there is no haze)—a line from a poem circulated on social media criticizing the CPC for attempting to clean up pollution in anticipation of foreign delegates arriving for APEC 2014 in Beijing, but not for citizens otherwise.

Thus, we see both collective action and government criticism clearly reflected in our unbiased keyword lists. While King *et al.* observed tolerance of government criticism on the platforms they studied, our analysis conclusively shows that SVPs and chat clients specifically target this kind of speech.

This paper provides a mapping of information controls across industry segments and raises questions for the research community. Diversity of tactics in implementing censorship undoubtedly lead to a diversity in what content is ultimately restricted. Furthermore, new government initiatives like "Clean the Web 2014" and the ongoing anti-rumor campaign [28] may quickly render outmoded previously confirmed theories. We thus offer a cautious note about applying any comprehensive theory about an ecosystem as varied and fast changing as the Chinese Internet.

Acknowledgments

We are grateful to Jakub Dalek for research assistance and Mamatjan Juma, Greg Fay, and Zubayra Shamseden for translation. Special thanks to Greg Wiseman for development work on the project website (https://china-chats.net). Jeffrey Knockel and Jason Q. Ng were supported by the Open Technology Fund Information Controls Fellowship Program. The research of the Citizen Lab was supported by the John D. and Catherine T. MacArthur Foundation.

References

[1] GuaGua, 法律声明. Available at http://www.guagua.cn/other/1907.html.

[2] Tian Ge Interactive Holdings, Global Offering. Available at http://www.tiange.com/Upload/Pigeon-Cover-IPO-ENG-2d.pdf.

[3] Tian Ge Interactive Holdings, Interim Report 2014. Available at http://www.tiange.com/Upload/e101.pdf.

[4] Tian Ge Interactive Holdings, Tian Ge Announces 2014 Third Quarter and Interim Results. Available at http://www.tiange.com/Upload/TIAN%20GE%20ANNOUNCES%202014%20THIRD%20QUARTER%20RESULTS.pdf.

[5] University of Hong Kong, WeiboScope. Available at http://weiboscope.jmsc.hku.hk/.

[6] YY Inc, 2014 Annual Report. Available at http://investors.yy.com/annuals.cfm.

[7] YY Inc, YY主播违规管理方法. Available at http://www.yy.com/1309/242131808402.html.

[8] Information Office of the State Council of the People's Republic of China. Available at http://english.gov.cn/2010-06/08/content_1622956.htm, 2010.

[9] Guagua: Exploring China's Online Video Community 'Goldmine'. *Knowledge@Wharton* (Oct. 2013). Available at http://knowledge.wharton.upenn.edu/article/guagua-exploring-chinas-online-video-community-goldmine/.

[10] China Digital Times, Steamed Bun Xi. Available at http://chinadigitaltimes.net/space/Steamed_Bun_Xi, 2014.

[11] China Launches Campaign to Cleanse Web of Terror Content. Available at http://www.reuters.com/article/2014/06/20/us-china-internet-xinjiang-idUSKBN0EV0TP20140620, June 2014.

[12] Towards a comprehensive picture of the great firewall's dns censorship. In *4th USENIX Workshop on Free and Open Communications on the Internet (FOCI 14)* (San Diego, CA, Aug. 2014), USENIX Association.

[13] Bamman, D., O'Connor, B., and Smith, N. A. Censorship and deletion practices in Chinese social media. *First Monday 17*, 3 (2012).

[14] Branigan, T. Chinese former security tsar Zhou Yongkang charged in corruption case. Available at http://www.theguardian.com/world/2015/apr/03/china-security-tsar-zhou-yongkang-charged-corruption., April 2015.

[15] Clayton, R., Murdoch, S. J., and Watson, R. N. M. Ignoring the Great Firewall of China. In *6th Workshop on Privacy Enhancing Technologies* (2006).

[16] Crandall, J., Crete-Nishihata, M., Knockel, J., McKune, S., Senft, A., Tseng, D., and Wiseman, G. Chat program censorship and surveillance in China: Tracking TOM-Skype and Sina UC. *First Monday 18*, 7 (2013).

[17] Crandall, J. R., Zinn, D., Byrd, M., Barr, E., and East, R. ConceptDoppler: A weather tracker for Internet censorship. In *14th ACM Conference on Computer and Communications Security, Oct.29-Nov2, 2007* (2007), pp. 1–18.

[18] Crete-Nishihata, M., Dalek, J., Hardy, S., Ng, J. Q., and Senft, A. Asia Chats: LINE Censored Keywords Update. Available at https://citizenlab.org/2014/04/line-censored-keywords-update/, 2014.

[19] Ensafi, R., Winter, P., Mueen, A., and Crandall, J. R. Analyzing the Great Firewall of China over space and time. In *Privacy Enhancing Technologies Symposium* (2015), De Gruyter Open.

[20] Hardy, S. Asia Chats: Investigating Regionally-based Keyword Censorship in LINE. Available at https://citizenlab.org/2013/11/asia-chats-investigating-regionally-based-keyword-censorship-line/, 2013.

[21] King, G., Pan, J., and Roberts, M. E. How censorship in china allows government criticism but silences collective expression. *American Political Science Review 107* (2013), 1–18.

[22] King, G., Pan, J., and Roberts, M. E. Reverse-engineering censorship in china: Randomized experimentation and participant observation. *Science 345* (2014), 1–10.

[23] Knockel, J., Crandall, J. R., and Saia, J. Three researchers, five conjectures: An empirical analysis of tom-skype censorship and surveillance. In *FOCI 2011: Proceedings of the USENIX Workshop on Free and Open Communications on the Internet* (2011).

[24] Link, P. China: The Anaconda in the Chandelier. *The New York Review of Books* (2002).

[25] MacKinnon, R. China's Censorship 2.0: How companies censor bloggers. *First Monday; Volume 14, Number 2 - 2 February 2009* (2009).

[26] MacKinnon, R. China's "networked authoritarianism". *Journal of Democracy 22*, 2 (2011), 32–46.

[27] Ng, J. Q. Tracing the Path of a Censored Weibo Post and Compiling Keywords that Trigger Automatic Review. Available at https://citizenlab.org/2014/11/tracing-path-censored-weibo-post-compiling-keywords-trigger-automatic-review/, 2014.

[28] Ng, J. Q. Politics, Rumors, and Ambiguity: Tracking Censorship on WeChat's Public Accounts Platform. Available at https://citizenlab.org/2015/07/tracking-censorship-on-wechat-public-accounts-platform/., July 2015.

[29] Park, J. C., and Crandall, J. R. Empirical study of a national-scale distributed intrusion detection system: Backbone-level filtering of HTML responses in China. In *30th International Conference on Distributed Computing Systems (ICDCS) 2010* (2010), pp. 1–12.

[30] Tang, A. China's 'evil' church demolition campaign continues, say activists. Available at http://www.telegraph.co.uk/news/worldnews/asia/china/11537435/Chinas-evil-church-demolition-campaign-continues-say-activists.html., April 2015.

[31] Villeneuve, N. Search monitor project: Toward a measure of transparency. Available at http://citizenlab.org/wp-content/uploads/2011/08/nartv-searchmonitor.pdf/, 2008.

[32] Villeneuve, N. Breaching Trust: An analysis of surveillance and security practices on China's TOM-Skype platform. Available at http://www.infowar-monitor.net/breachingtrust/, 2009.

[33] wa Fu, K., hong Chan, C., and Chau, M. Assessing censorship on microblogs in china: Discriminatory keyword analysis and the real-name registration policy. *IEEE Internet Computing 17*, 3 (2013), 42–50.

[34] Weaver, N., Sommer, R., and Paxson, V. Detecting forged tcp reset packets. In *NDSS* (2009), The Internet Society.

[35] Winter, P., and Lindskog, S. How the Great Firewall of China is Blocking Tor. In *Free and Open Communications on the Internet* (Bellevue, WA, USA, 2012), USENIX Association.

[36] Zhu, T., Phipps, D., Pridgen, A., Crandall, J. R., and Wallach, D. S. The velocity of censorship: High-fidelity detection of microblog post deletions. In *Presented as part of the 22nd USENIX Security Symposium (USENIX Security 13)* (Washington, D.C., 2013), USENIX, pp. 227–240.

[37] Zittrain, J., and Edelman, B. Internet filtering in China. *Internet Computing 7*, 2 (2003), 70–77.

Notes

[1]The raw and processed dataset with visualizations is available at https://china-chats.net