(CAN'T) PICTURE THIS 2

An Analysis of WeChat's Realtime Image Filtering in Chats

By Jeffrey Knockel and Ruohan Xiong

JULY 15, 2019

RESEARCH REPORT #122

munk school OF GLOBAL AFFAIRS & PUBLIC POLICY

UNIVERSITY OF TORONTO

THE CITIZEN LAB

Copyright

Document Version: 1.0

Suggested Citation

Jeffrey Knockel and Ruohan Xiong. "(Can't) Picture This 2: An Analysis of WeChat's Realtime Image Filtering in Chats," Citizen Lab Research Report No. 122, University of Toronto, July 2019.

Acknowledgements

This project was supported by Open Society Foundations. We would like to thank an anonymous researcher for assistance in translating and categorizing the images featured in this report. We would also like to thank Ron Deibert for advising this work, Simon Humbert for technical assistance, as well as Masashi Crete-Nishihata, Miles Kenyon, and Adam Senft for helpful suggestions and peer review.

The full image data set that we collected and the codebook we used to categorize each image are available here.

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

The Citizen Lab is an interdisciplinary laboratory based at the Munk School of Global Affairs and 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.

Contents

点击这里阅读中文报告简要

Key Findings

- **WeChat implements realtime, automatic censorship of chat images based on text contained in images and on an image’s visual similarity to those on a blacklist**
- **WeChat facilitates realtime filtering by maintaining a hash index populated by MD5 hashes of images sent by users of the chat platform**
- **We compare levels of filtering across WeChat’s Moments, group chat, and 1-to-1 chat features and find that each has different images censored; we find that Moments and group chat are generally more heavily filtered than 1-to-1**
- **WeChat targets predominantly political content including images pertaining to government and social resistance**
- **WeChat’s image censorship is reactive to news events; we found censored images covering a wide range of events, including the arrest of Huawei’s CFO, the Sino-US Trade War, and the 2018 US Midterm Elections**

Summary

Internet platform companies operating in China are required by law to control content on their platforms or face penalties, under the expectation that companies will invest in the technology and personnel required to ensure compliance. These requirements form a system of intermediary liability or “self-discipline” in which Internet platform companies are held liable for content on their services. Previous work has found little consistency in what content different Chinese Internet platforms censor. However, some high profile Internet platforms are known to frequently receive government directives.

In this report, we study how Tencent, one of China’s largest Internet companies, implements image filtering on WeChat. WeChat is China’s most popular social media platform with over one billion monthly active users. Its functionality includes 1-to-1 and group chat, Moments (a feature for posting messages and images similar to Facebook’s Timeline), and other social networking features. In a previous report, we studied automatic image filtering on WeChat Moments and found that it employed lengthy, computationally expensive operations to determine if an image is sensitive. However, in this report, we focus on WeChat’s chat functionality, where image filtering, if it is to exist, must operate in realtime.

We found that Tencent implements realtime, automatic censorship of chat images on WeChat based on text contained in images and on an image's visual similarity to those on a blacklist. Tencent facilitates realtime filtering by maintaining a hash index populated by MD5 hashes of images sent by users of the chat platform. If the MD5 hash of an image sent over the chat platform is not in the hash index, then the image is not filtered. Instead, it is queued for automatic analysis. If it is found to be sensitive, then its MD5 hash is added to the hash index, and it will be filtered the next time a user attempts to send an image with the same hash.

This finding indicates that censorship measurement—like the kind conducted in this report—not only evaluates censorship but can also influence and modify the behaviour of a realtime, automatic censorship system by introducing novel items that can be flagged as sensitive and subsequently censored. This helps us understand previous measurements and has implications for future censorship measurement research.

With an accurate understanding of how chat image filtering works on WeChat, we compare levels of filtering across Moments, group chat, and 1-to-1 chat. We find that each feature has different images censored, where Moments and group chat are generally more heavily filtered than 1-to-1 chat. We also present the first broad content analysis of images censored by Tencent, based on a set of 220 filtered images we discovered and categorized, many of which are event-related or critical of the Chinese government. Popularly referenced topics include the arrest of Huawei CFO Meng Wanzhou, the Sino-US Trade War, and the 2018 US Midterm Elections.

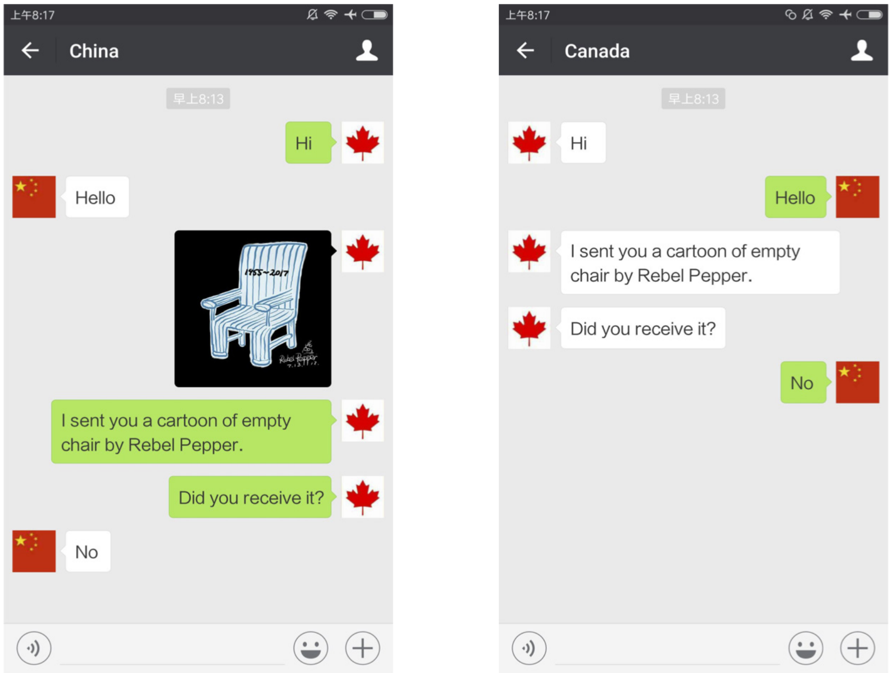
Previous work

In previous work, we found that when a message, post, or image is filtered on either WeChat chat features or WeChat Moments, it is hidden from the view of other users whose accounts are registered to mainland Chinese phone numbers, but remains visible to those registered using international phone numbers. This finding means that content filtering on WeChat is non-transparent, as no explicit notice is given to users when content is filtered and the original content remains visible to the user that sent or posted it.

In the predecessor to this report, we extensively analyzed how images are censored on WeChat's Moments platform. We found that after images are posted, they are censored using two different techniques: an Optical Character Recognition (OCR)-

based method that compares text in the image to keywords on a sensitive keyword blacklist and a visual-based method that compares the posted image's visual fingerprint to those on a sensitive image blacklist. If an image is deemed sensitive through either of these methods, then it is made invisible to all users with accounts registered to mainland Chinese phone numbers, except the original poster of the image. We found that this process is computationally expensive and that it often takes multiple seconds to censor an image after it has been posted.

While in that report we studied image censorship of images posted to WeChat Moments, that study did not analyze filtering of images sent in realtime through 1-to-1 chat or group chat. In this report, we study how WeChat identifies sensitive images sent over chat in realtime, despite the computationally expensive demands of detecting an image's blacklisted text or its visual similarity to that of a blacklisted image. We also expand upon the previous report by providing the first broad analysis of images censored by WeChat.



Hi

Hello

I sent you a cartoon of empty chair by Rebel Pepper.

Did you receive it?

No

Hi

Hello

I sent you a cartoon of empty chair by Rebel Pepper.

Did you receive it?

No

Figure 1: Left, a Canadian account sending an image memorializing Liu Xiaobo over 1-to-1 chat; right, the Chinese account does not receive it.

In an earlier report studying WeChat censorship of content related to the "709 Crackdown" on Chinese human rights defenders, we presented evidence of the existence of censorship of images sent over WeChat's chat functions. In a later report written in the wake of Liu Xiaobo's death, we measured the filtering of images related to Liu Xiaobo across WeChat's Moments, group chat, and 1-to-1 chat functions (see Figure 1), finding that automatic filtering of such content occurred across all three functionalities. We found inconsistent results measuring filtering over the chat functions, finding little filtering over chat one day but much more the next. We also found that Moments had a much higher level of image filtering than group chat. In this report, we explain how these two findings were artifacts created as a product of how WeChat implements realtime image filtering over chat and the research methodology that we used at the time. In doing so, we provide for the first time an accurate measurement of the level of filtering on Moments, group chat, and 1-to-1 chat.

Research Findings

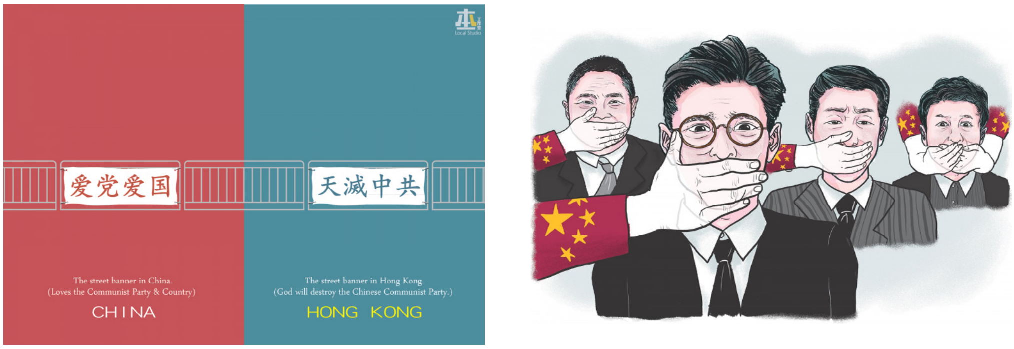

爱党爱国
天滅中共
The street banner in China.
(Loves the Communist Party & Country)
CHINA
The street banner in Hong Kong.
(God will destroy the Chinese Communist Party.)
HONG KONG

Figure 2: Left, an image blocked via OCR methods due to containing blacklisted text (天滅中共); right, an image blocked due to its visual similarity to a blacklisted image.

In this section, we explain our findings from analyzing the filtering mechanisms used to perform realtime image filtering on WeChat's chat features. We found that WeChat identifies sensitive images to censor over 1-to-1 and group chat features using the same two methods we had previously found to be used to censor images posted to WeChat Moments; namely, an OCR-based and a visual-based method. Figure 2 shows examples of images which were filtered based on the two respective methods. Both of these methods are computationally expensive and cannot be performed in realtime. We document these general methods in greater detail in the previous report.

WeChat populates a hash index using images users send over chat

The OCR-based and visual-based algorithms we discovered being used to filter images on WeChat Moments are too computationally expensive to be applied to realtime filtering of chat. In order to filter images in realtime, we found that WeChat uses another data structure called a *hash index*[1]. When a user sends an image, upon receipt by one of WeChat's servers, the server calculates its cryptographic hash and if the hash is in the hash index, then the image is filtered in realtime instead of being relayed to the intended user.

Cryptographic hashing is a technique used to quickly map the data contained in a file to a fingerprint, or hash, which is of fixed length and is ideally unique to the file analysed. Such hashes are designed so that the smallest of changes to the hashed file have, on average, as large of changes to the resulting hash as large changes to the hashed file. A change in hashes can confirm that a file has been modified but not by how much the file has been modified. Cryptographic hashes can be computed quickly, and therefore this hashing is amenable to realtime filtering applications unlike more expensive techniques such as OCR or perceptual fingerprints. However, cryptographic hashes are highly inflexible, as even small changes to an image or its metadata radically alter its cryptographic hash.

To overcome these limitations, we found that WeChat still employs the more computationally expensive OCR-based and visual-based techniques to populate the hash index in non-realtime. We observed this by making small changes to the file metadata of a blacklisted image, which makes no changes to the appearance of the image but still makes substantial changes to its cryptographic hash. Whenever we modified an image in this manner, it would be received unfiltered the first time it was sent. However, when we sent the same modified image again seconds later, it would be filtered. Any modification to the sent image file, including to its metadata, would allow it to evade filtering once, as any such modification would change the file's hash to one not yet present in the hash index.

Group chat and 1-to-1 chat have different indexes

In previous reports, we found that WeChat uses a different blacklist for filtering group chat—a chat featuring more than two users—as opposed to 1-to-1 chat—a conversation between two users. Similarly, although both 1-to-1 and group chat

[1] In this report, we use *image blacklist* to refer to the list of blacklisted images to which an image sent over chat is visually compared in order to determine if it is sensitive. We use *keyword blacklist* to refer to the list of keyword combinations that cause an image to be filtered if the image is determined to contain one of these keyword combinations via OCR. We use *hash index* to refer to the index of MD5 hashes that Tencent uses to filter images in realtime chat based on whether those images were previously found to be on the image blacklist. We use *sensitive image* to refer to an image that, if sent, would be added to the hash index via either one of the visual-based or OCR-based methods.

used hash indexes, they did not share the same index. Taking an image that is blacklisted on 1-to-1 chat, group chat, and Moments, and modifying the file so that it has a unique cryptographic hash, we made the following observations:

1) Sending that image file once over group chat does not cause images with that hash to be filtered on 1-to-1 chat until you also send it over 1-to-1 chat.
2) Similarly, sending that image file once over 1-to-1 chat does not cause images with that hash to be filtered on group chat until you also send it over group chat.
3) Sending an image file over Moments does not cause images with that hash to be filtered on group chat or 1-to-1 chat.

These observations indicate that the hash indexes used by 1-to-1 chat and group chat are independent. This result is surprising, since if an image is blacklisted on both functionalities, then there is no reason for an image's hash not to be added to both functionalities' indexes if it is found sensitive on one of them. Similarly, even though we did not find Moments to use a hash index, when a sensitive image is filtered, if it is also blacklisted on 1-to-1 chat or group chat, it would seem worthwhile to perform an inexpensive cryptographic hash of the image file and add it to those hash indexes.

Choice of WeChat client and settings affect user experience of filtering

When sending images over WeChat's chat features, we found that the choice of WeChat client and client settings could affect whether an image is filtered. For instance, if an image file is sent over chat using WeChat's Windows client, then the outcome of whether the image is filtered may be different than if that image had been sent using the Android client.

To explain these inconsistencies, we analyzed the behaviour of both the Windows and Android WeChat clients to determine how they handle image uploads. We found that the Windows version performs no modifications to an image before it is uploaded and uploads the original image file. Alternatively, we found that the behaviour of the Android client is partially dependent on how the image is uploaded:

- If "Full Image" ("发送原图") is not selected, then the image is sometimes reencoded. The necessary conditions for image re-encoding are not clear but may depend on the original image's size.

- The Android client does not re-encode if you select "Full Image."

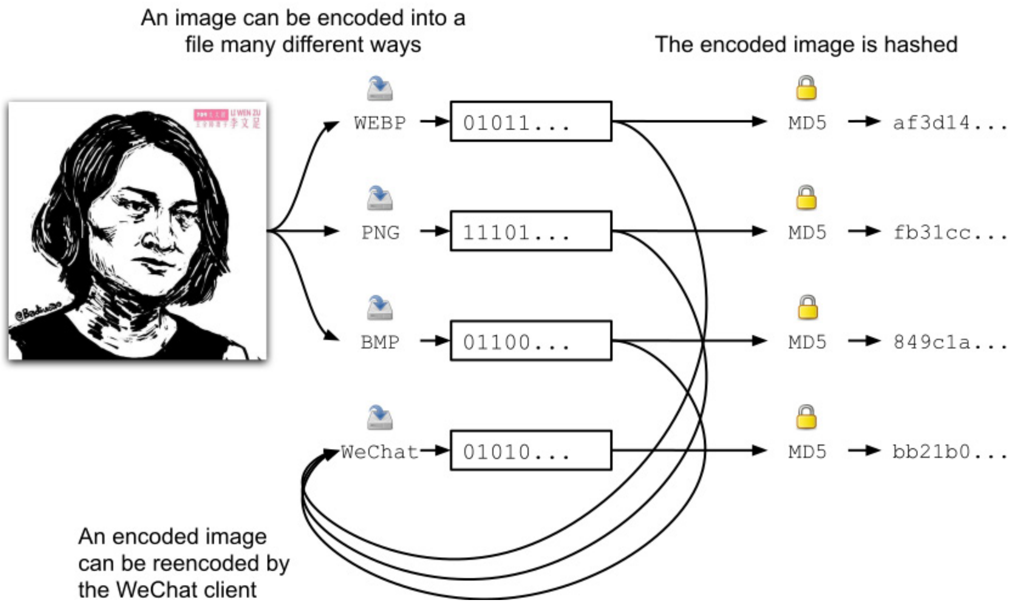

An image can be encoded into a file many different ways
The encoded image is hashed
WEBP
01011...
MD5
af3d14...
PNG
11101...
MD5
fb31cc...
BMP
01100...
MD5
849c1a...
WeChat
01010...
MD5
bb21b0...
An encoded image can be reencoded by the WeChat client

Figure 3: There are many different ways to encode an identically appearing image, each of which have different hashes. However, if the image is re-encoded by the WeChat client, then any identically appearing image will have the same hash.

We found that the use of client re-encoding can make the image filtering more powerful in that it effectively results in an image's hash representing all images with identical pixel values as opposed to merely a specific image encoding. The process of re-encoding extends the generalizability of hash-based filtering because any image containing the same pixel contents will be encoded to the same file and thus have an identical hash (see Figure 3). When this happens, any changes to the original image's encoding or to its metadata will be ineffective at evading filtering, and some change to the image's pixel values, if even a small one, will be required to change the resultant hash. When the client does not re-encode, then any change to an image's file encoding, including to its metadata, is sufficient to change its hash.

Group chat (and Moments) blacklist more images than 1-to-1

With an improved understanding of how image filtering happens over WeChat's chat platforms, we wanted to revisit the question of whether the platform had increased filtering on different features depending on the number of users the content has the potential of reaching.

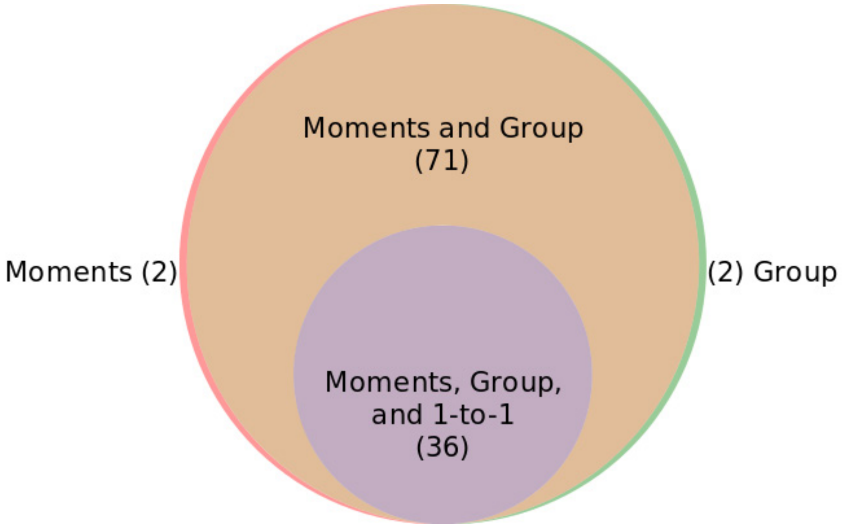


Figure 4: Out of 111 images, the # of images censored on each WeChat feature; red is only Moments, green is only group chat, brown is both Moments and group chat, and purple is Moments, group chat, and one-to-one chat.

On November 15, we tested images previously found censored on WeChat during the 709 Crackdown and on the days surrounding Liu Xiaobo's passing. Together this dataset comprised 128 images. We found that 111 of these images were still censored on some feature on WeChat. We found that 36 out of 111 images were filtered in 1-to-1 chat, and each of these were also filtered in group chat and on Moments (see Figure 4). We also found that 107 of the 111 images were filtered on both group chat and Moments, with 2 images being filtered only on group chat and another 2 images being filtered only on Moments. This finding suggests that, while the image blacklists for group chat and Moments are largely identical, the reduced filtering on 1-to-1 chat is an indication that WeChat perceives 1-to-1 chat as less of a risk for sensitive conversations due to its more private nature.

This result contrasts with our observations of image filtering during two days of testing around Liu Xiaobo's passing, when we found that fewer images were filtered on group chat than Moments. This earlier result was likely an artifact of how image filtering was tested. In this earlier work, since many images were only tested once on the last day of testing, the testing only occurred long enough to insert their hashes into the hash index but not to observe filtering. In fact, across the two days of testing, many images that were tested on the first day which were not filtered initially were found to be filtered when retested on the second day.

WeChat uses MD5 hashes

Thus far, due to the observed properties of WeChat's hash index, we suspected that they were using a cryptographic hash to index images, but we did not yet know which hash specifically. Due to its mention in this Tencent-owned patent that we discovered in a previous report and due to its general popularity as a file hash, we suspected that WeChat may be using the MD5 hashing algorithm to hash images. As a result of vulnerabilities in the MD5 hash function, we were able to test whether this is actually the case.

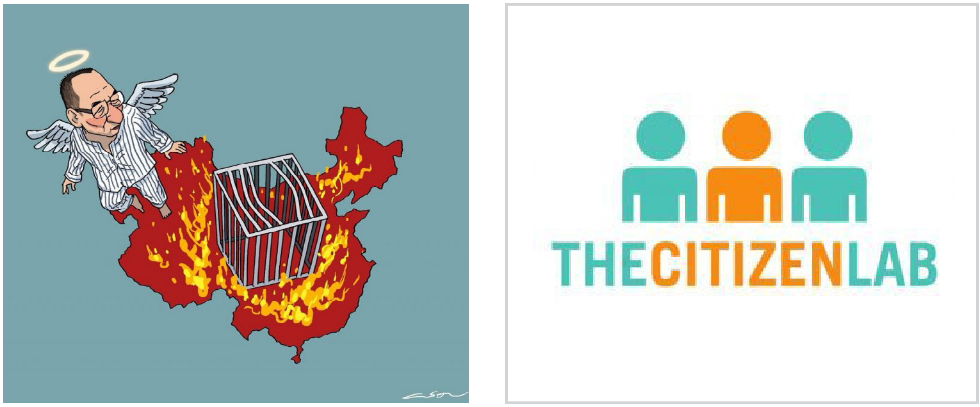


Figure 5: Left, a picture memorializing Liu Xiaobo; right, the logo of the Citizen Lab. Using a chosen-prefix collision attack, we modified the contents of these files such that both of these images have the same MD5 hash.

As a cryptographic hash function, MD5 is well known to be broken in regards to collision resistance and is vulnerable to chosen-prefix collision attacks, a type of attack where specially calculated data blocks are appended to the data of two arbitrary files to generate a pair of modified files which hash identically. Like many image formats, the JPEG standard defines an end of image marker. During the rendering process, all file content after this marker is ignored, meaning that the appended data does not affect the appearance of the image. This behaviour allowed us to design an experiment testing whether Tencent used MD5 to hash images sent over chat. Using Project HashClash, we generated two JPEGs which shared the same MD5 hash, one a blacklisted image of Liu Xiaobo and another of the Citizen Lab logo, an image which is not normally filtered (see Figure 5).[2] This operation took about five hours of computation time on an Amazon Web Services (AWS) g3.4xlarge

[2] Since our initial generation of a chosen-prefix collision, a new exploit for creating MD5 hash collisions in specific image formats has been published which can be performed computationally instantly. Using the same two JPEGs, we generated another collision using this method and repeated our experiment, finding the same results.

GPU-optimized instance. After sending the blacklisted image, we sent the Citizen Lab logo which we generated to have the same MD5 hash. The Citizen Lab logo was also filtered, confirming that the hash index uses MD5 as its hash function.

At the additional cost of space, WeChat could maintain a hash index of non-sensitive image hashes, in addition to having a hash index of sensitive image hashes. The use of an additional index would save computational resources, as WeChat would then not have to determine whether any non-sensitive image is sensitive once its hash has been added to the non-sensitive hash index. To test whether WeChat uses a non-sensitive hash index, we generated another collision between a blacklisted image of Liu Xiaobo and of the Citizen Lab logo, and this time we sent the images in the opposite order: the Citizen Lab logo first followed by the Liu Xiaobo image. We then sent the Liu Xiaobo image a second time to test whether it was filtered. We found that sending the Citizen Lab logo first did not prevent the Liu Xiaobo image from being filtered. This result shows that there is no non-sensitive hash index that preempts checking a sent image for sensitivity. One possibility why such a system is not implemented is that a non-sensitive index would need to be invalidated each time WeChat adds a new image to the image blacklist, as hashes previously determined to be not sensitive could have become sensitive.

Finally, we wanted to test whether sending a non-sensitive image file would remove that file’s hash from the sensitive hash index. To test this, we sent the now-filtered Citizen Lab logo a second time to test whether sending it the first time removed it from the hash index. We found that the image was filtered both times. Thus, sending a non-sensitive image file does not remove its hash from the hash index if it is present.

The MD5 algorithm’s vulnerability to chosen-prefix collision attacks allows a user to trick WeChat’s image filtering system to filter non-sensitive images by having their hashes collide with those of blacklisted images. However, these images must be under the user’s control and have already been specifically modified by the user, since the MD5 collision attack analyzes and modifies both the sensitive image file and the non-sensitive image file simultaneously. Since image files already being sent over the platform would not generally have been modified in the specific way necessary to create the collision, this attack could not be used to cause popular images already being sent over the platform to be filtered.

Hash index eviction

In this section, we explore if and how Tencent decides to remove hashes from their hash index. In our testing, we had anecdotally observed image files which had been added to the hash index to be no longer be in the hash index when tested later. We sought to rigorously measure the hash eviction behaviour of Tencent's hash index mechanism.

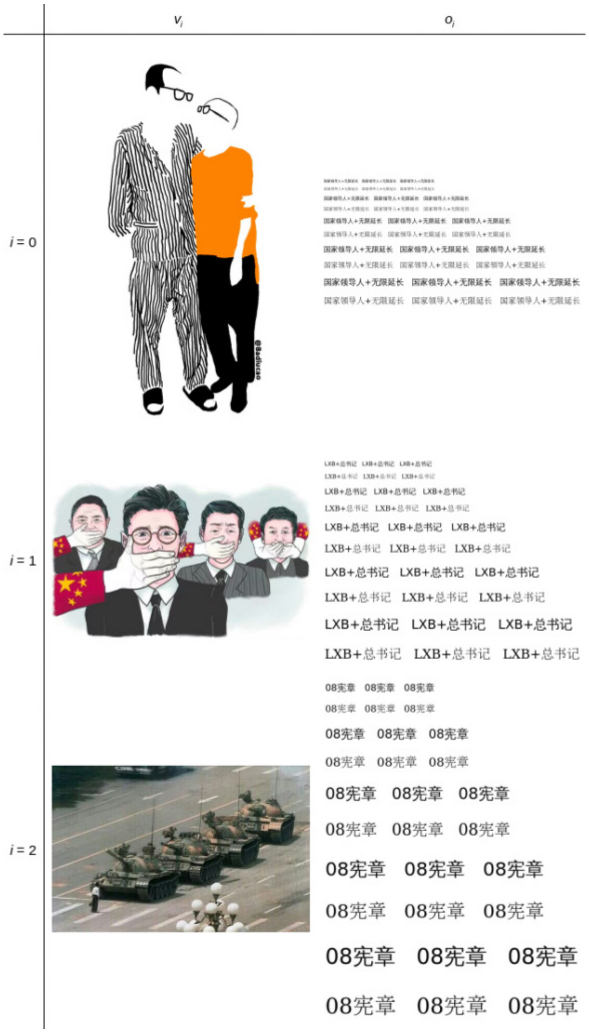


Figure 6: The six different source images tested.

We measured hash eviction using six different source images comprising three *test pairs*, as shown in Figure 6. Each test pair *i*, 0 ≤ *i* < 3, consisted of one source image *vi* filtered by visual similarity and one source image *oi* filtered by OCR. Each image filtered by visual similarity were politically sensitive images: one concerning Liu Xiaobo's death, one related to the 709 Crackdown, and one of the historical picture of "Tank Man." Each image filtered by OCR were images generated containing a politically sensitive keyword combination known to be filtered. The three sensitive keyword combinations we used were “国家领导人 [+] 无限延长” (national leader [+] unlimited extension), “LXB [+] 总书记” (Liu Xiaobo [+] General Secretary), and “08宪章” (Charter 08). To generate an image from each of these phrases, we created an image containing the keyword combination multiple times in a variety of font sizes and styles. Specifically, we created an image with ten lines each containing the keyword combination repeated three times, where the font size of line ℓ, $0 \leq \ell < 10$, is $12 + 4 \cdot \lfloor \ell / 2 \rfloor$ and where the even lines are written in a 黑 (Hei) typeface and the odd lines are written in a 明体 (Ming) typeface.

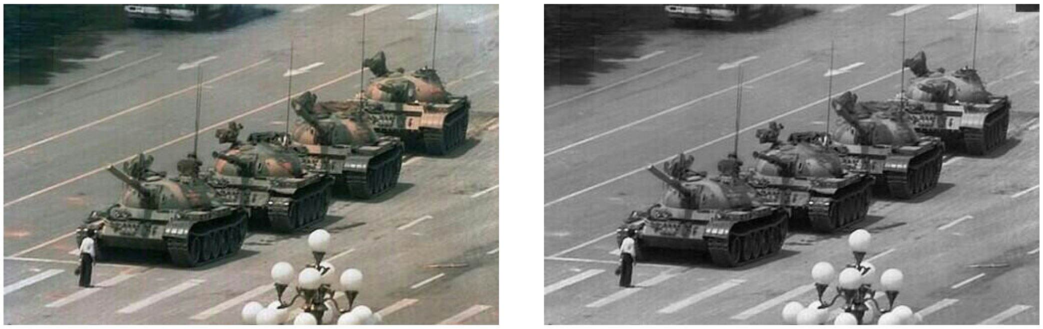


Figure 7: Left, the original image v_2; right, a canary modified to be grayscale and to have three blocks of modified pixels in the upper-right corner.

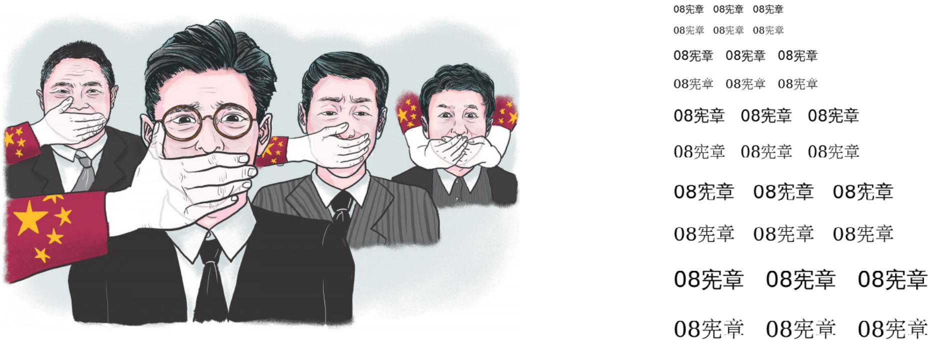
For each visually filtered source image v_i and each OCR filtered source image o_i, we created 33 subtle variations we call *canaries*. We use these to measure the lifespan of each image’s hashes in the hash index. When a canary “dies” (i.e., the hash index is no longer filtered), then we know that that hash index has been evicted.

For a given source image, we create canary d, $0 \leq d < 33$, as follows. We first convert the canary’s source image to grayscale. Then we modify a region of pixels in the image such that, for each grayscale pixel value g in this region, we replace it with value $(g + 128) \bmod 256$, where 256 is the maximum value of the grayscale channel, effectively changing each pixel value by half of its intensity range. The region of pixels to modify in this manner is a rectangle whose dimensions and location are determined as a function of d. Vertically, the region always has a height of 8 pixels and begins at the top edge of the image extending downward. Horizontally, the region has a width of $8 \cdot (d + 1)$ pixels and begins at the right edge of the image extending leftward. Figure 7 shows an example of the modifications we made to create one canary.

To test eviction behaviour, for each test pair i, we primed WeChat’s hash index on November $15 + i$ with visually filtered source image v_i and OCR filtered source image

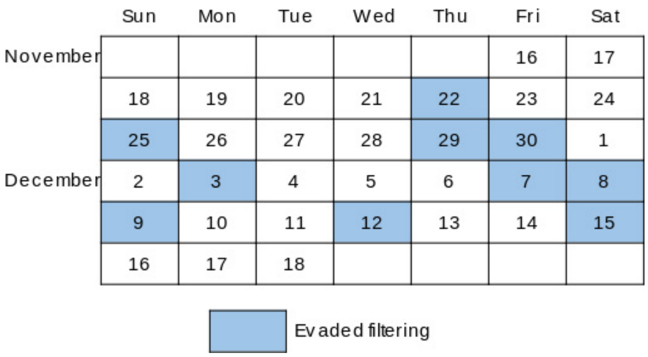
o_i . To prime the cache, we first sent each canary of v_i and o_i. In each case we found that they were never filtered. Minutes afterward we resent each canary to confirm that the canaries' hashes were now in WeChat's hash index. We found that in each case they were filtered as they were now in the hash index.

To test how long each canary survived in the hash index, each day we formed a test for the eviction of a canary. For each test pair *i*, we again sent its canary *d* in the afternoon of the day *d* days after the images were originally planted in the hash index (November 15 + *i* + *d*) to test if the canary had survived in the index *d* days. In other words, for each source image we tested whether one of its canaries was evicted each day up to a month later.



08宪章 08宪章 08宪章
08宪章 08宪章 08宪章
08宪章 08宪章 08宪章
08宪章 08宪章 08宪章
08宪章 08宪章 08宪章
08宪章 08宪章 08宪章
08宪章 08宪章 08宪章
08宪章 08宪章 08宪章
08宪章 08宪章 08宪章
08宪章 08宪章 08宪章

Figure 8: Left, v_1*, an image depicting Hong Kong censorship; right,* o_2*, an image referring to Charter 08, a human rights manifesto; some canaries of each of these images evaded filtering.*



	Sun	Mon	Tue	Wed	Thu	Fri	Sat
November						16	17
	18	19	20	21	🟦 22	23	24
	🟦 25	26	27	28	🟦 29	🟦 30	1
December	2	🟦 3	4	5	6	🟦 7	🟦 8
	🟦 9	10	11	🟦 12	13	14	🟦 15
	16	17	18				

🟦 Evaded filtering

Figure 9: Days between November 16 and December 18 when image v_1*'s canary image evaded filtering.*

We found that every canary except ten of v_1 and one of o_2 were filtered (see Figure 8). Only one canary of o_2, the one tested on December 10, evaded filtering. The ten canaries that evaded filtering of v_1 are as shown in Figure 9.

To test eviction up to a second month, we performed a second experiment by reusing our canaries which were replanted or refreshed during our previous experiment. For each test pair *i*, we tested all 33 of each of v_i and o_i's canaries in the evening of January 16 + *i*. Note that the test date is not a function of *d* as the canaries were already replanted or refreshed each day for a period of one month. Thus, each source image's canaries must be tested on the same day. In this experiment, we found that all of these canaries were filtered (i.e., that none evaded filtering).

At a high level, we discovered a general trend that there is a higher chance of an image being evicted as time goes on. However, the results were largely nondeterministic, and it would be unreliable to depend on image hash eviction behaviour to evade image filtering. In future work, testing eviction over a longer test period may possibly produce more clear cut results.

One observation we found is that eviction of a hash appears to be a function of the blacklisted image to which it corresponds, which we observed in only v_1 experiencing a large amount of eviction. This observation may be a result of each blacklisted image having its own independent limit for the number of hashes allowed in the index. If this hypothesis is true, then this would suggest that v_1 may have been more frequently shared on the platform at the time of testing.

Blacklist removal

In some cases, we observed that an image had been removed from WeChat's blacklist but its MD5 hashes nevertheless appeared to remain in the hash index. For instance, on May 30, 2019, we observed this behaviour with the iconic "Tank Man" photo (see Figure 7) in 1-to-1 chat. If we modified the metadata in the image file, then the photo was never filtered. However, if we sent the image unmodified, then it was. This inconsistency suggests that hashes of the image were in the hash index but that the image was no longer blacklisted and that new hashes were no longer being added to the index. Thus, it would seem that when images are removed from the blacklist, their hashes are not immediately removed from the hash index.

Summary of research findings

We found that WeChat uses OCR- and visual-based filtering to identify sensitive images posted to Moments, group chat, and 1-to-1 chat. Filtering on Moments occurs in non-realtime after images are analyzed, as this is a computationally expensive process. To implement realtime filtering for images sent in chat, WeChat uses a hash index of known sensitive images and filters any images whose MD5 hash exists in the index. Separate hash indexes are maintained for group chat and 1-to-1 chat, and if an image is not found in the respective index, it is delivered to its intended recipient(s) but queued for later analysis for sensitivity by OCR- and visual-based methods. If an image is found to be sensitive, it is added to the hash index, preventing the future delivery of the same file. We found that known hashes of sensitive images can be removed from the index after some time; however, it is currently unclear how the index evicts hashes. Moreover, even after an image is no longer blacklisted, we found that hashes of the file can persist in the index.

Analyzing filtered image content

In this section, we aim to characterize what images are filtered and identify their topics. In previous work, we used automated testing to determine which keyword combinations are filtered on WeChat. While this approach is generalizable to image filtering, it resulted in WeChat banning our test accounts because it suspected them to be spam. Moreover, we found that new accounts required approval from a second account that must have existed for over six months, be in good standing, and have not already approved any other accounts in the past month. Because of these requirements, we found that creating new WeChat accounts was prohibitively difficult.

To automatically test for image filtering on WeChat, we investigated other Tencent-operated services that may be using the same or similar image blacklists as WeChat. Previous research has found positive but not complete correlation between companies and the content filtered across their products. We found that Tencent's Qzone (QQ空间), a blogging platform which has been the subject of previous censorship research, used the same two methods to filter images as WeChat—an OCR-based and a visual-based method—and that its blacklist appeared similar to that of WeChat's.

In the remainder of this section, we describe our method for identifying 220 filtered images on Qzone. We then characterize and analyze these 220 images. Finally, we select a random sample of these 220 filtered images and test whether they are also filtered on WeChat in order to evaluate how similar the image blacklists are between these two Tencent platforms.

Measuring filtered images

Our previous work analyzing image filtering on WeChat has focused on specific topics of censorship, such as Liu Xiaobo or human rights crackdowns. However, in this work we aimed to gain a broader understanding of which images Tencent automatically filters. In order to do this, we sampled content from a broad set of images likely to be filtered. To construct our sample, we utilized [WeChatscope](), a project that automatically records deleted posts on the WeChat Public Posts platform. We looked at posts WeChatscope detected as deleted within a six month period, between October 17, 2018 and April 17, 2019. We constructed our sample set to consist of every image from each of these deleted posts.

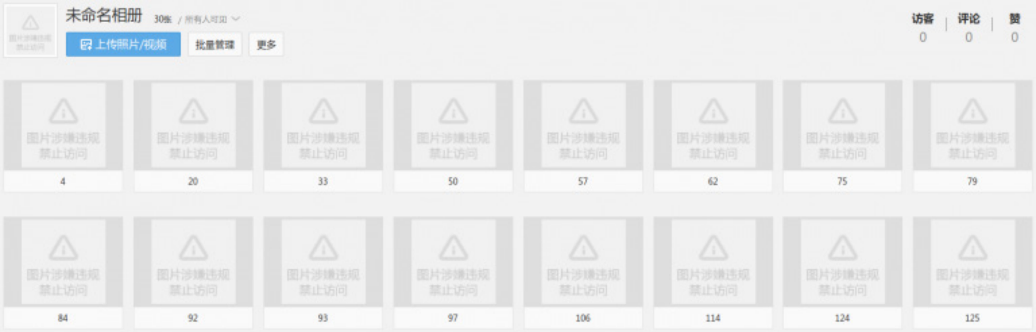


Figure 10: An album containing censored images on Qzone.

To test whether each image was automatically filtered, we wrote a Python script to automatically upload each image to Qzone. After uploading five images in succession, the program waits 60 seconds. It then checks to see if each image has been replaced with that of a placeholder image communicating that the original image has been deleted (see Figure 10), which we detect as any file beginning with the GIF magic number and containing fewer than 2048 bytes. We ran this script from March 14, 2019, to April 22, 2019. During this test period, the QQ account that we used for testing was never banned.

Analysis of filtered images

Using the technique described in the previous section, we found 220 filtered images. Eight of them had duplicated content, and so we excluded them from the analysis. We analyzed the remaining 212 images based on their content and the context of the original WeChat public account articles they originated from.

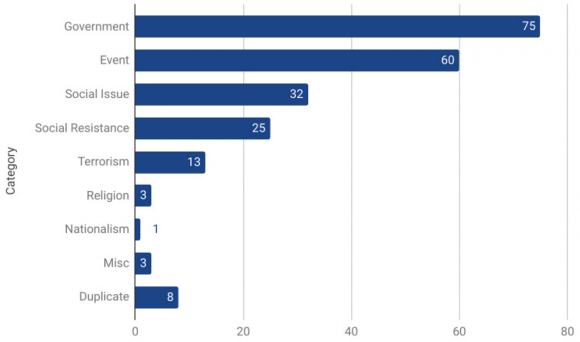


Figure 11: Distribution of censored images by category.

We coded each image into content categories based on a code book we developed for this report. The most popular content categories were Chinese Government (75 images) and Events (60 images). Figure 11 shows a full breakdown. In the remainder of this section, we present a curated list of examples from a number of these categories.

Government

We found 75 censored images with content related to the Chinese government. These include not only images critical of the government such as sarcastic cartoons (see Figure 12) but also neutral representations of government policy and photos of government leaders and party cadres.

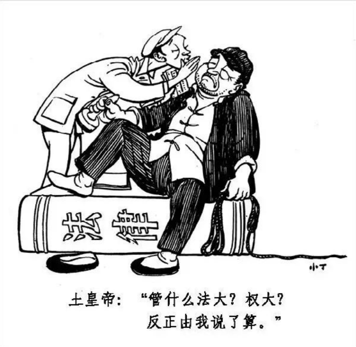


Figure 12: The caption in the cartoon reads, "Emperor: Who cares about whether law or power is more powerful? In the end, I am the one that dictates everything."

It is unclear how Tencent decides which images to filter. In previous research, we found that WeChat censored neutral keywords referencing official party policies and ideology around highly sensitive events. Similarly, this report finds that censored images of or alluding to government leaders and party cadre were not necessarily negative.

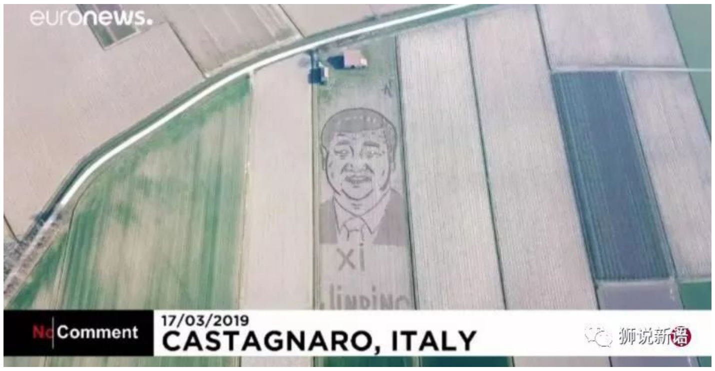


Figure 13: Screenshot of a European news program.

In one example (see Figure 13), we found that a screenshot captured from a news broadcast by Euronews, a European television network based in France, was filtered. The news clip is about an artist in Italy who used a tractor to create a huge portrait of Chinese President Xi Jinping smiling ahead of Xi's visit to Italy in March 2019.

Events

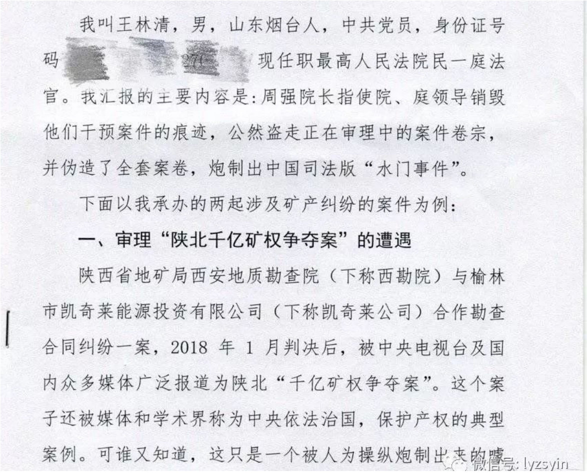
Chinese social media censorship is often reactive to news cycles, and companies tend to tighten their information controls around sensitive events. We found a total of 60 images that reference 10 distinct events: eight news events which happened around or close to our test period and two historical events.

Events	Note	# Images Censored
Cultural Revolution	A nationwide movement launched by Chairman Mao Zedong in 1966 to preserve communist values. The movement lasted until 1976.	4
1989 Tiananmen Movement	The Tiananmen Square protests in 1989 are a persistently taboo topic in China.	1
Fan Bingbing Tax Evasion Scandal	Fan Bingbing, one of China's highest-earning entertainers, was caught in a tax evasion scandal in 2018.	2
2018 Chongqing Bus Crash	On October 28, 2018, a bus plunged off the Second Wanzhou Yangtze River Bridge into the Yangtze River in Wanzhou District, Chongqing, causing at least 15 deaths.	2
2018 US Midterm Elections	US midterm elections were held in November 2018 featuring hundreds of congressional, state, and local seats in contest.	3
Supreme People's Court Scandal	In December 2018, the Supreme People's Court of China acknowledged the disappearance of court documents related to a contract dispute between two mining companies.	24
Huawei Saga	Huawei's CFO Meng Wanzhou was arrested in Canada under charges of fraud and violating international sanctions.	10
Sino-US Trade War	The so-called trade war between China and the United States engaged via increasing tariffs since 2018.	8
CRISPR-baby Scandal	Chinese scientist He Jiankui announced in February 2019 the birth of twin girls with edited genomes which he engineered. The announcement triggered international debates over research ethics.	2

2019 Chengdu Campus Food Scandal	In March 2019, mouldy bread and rotting meat were found at Chengdu No 7 Experimental High School. The incident caused protests among parents.	2
2019 Sichuan Forest Blaze	In late March 2019, a forest fire broke out in Sichuan. Thirty firefighters sent to tackle the fire were killed during the mission.	2

Table 1: Events referenced by images censored on Qzone, in chronological order. Some events above are still unfolding.

Descriptions of the events referred to by images in our dataset are listed in Table 1. A large number of event-related images in our dataset referenced scandals, the Sino-US Trade War, and the arrest of Huawei CFO Meng Wanzhou.



我叫王林清，男，山东烟台人，中共党员，身份证号码[redacted]现任职最高人民法院民一庭法官。我汇报的主要内容是：周强院长指使院、庭领导销毁他们干预案件的痕迹，公然盗走正在审理中的案件卷宗，并伪造了全套案卷，炮制出中国司法版“水门事件”。

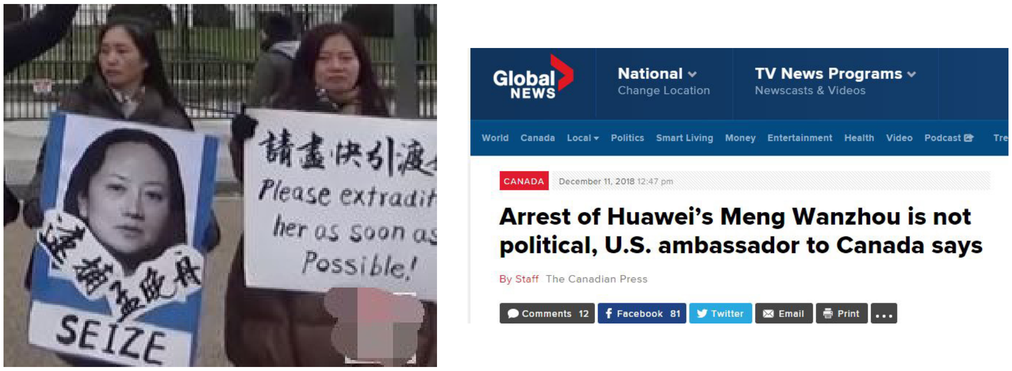
下面以我承办的两起涉及矿产纠纷的案件为例：

一、审理“陕北千亿矿权争夺案”的遭遇

陕西省地矿局西安地质勘查院（下称西勘院）与榆林市凯奇莱能源投资有限公司（下称凯奇莱公司）合作勘查合同纠纷一案，2018 年 1 月判决后，被中央电视台及国内众多媒体广泛报道为陕北“千亿矿权争夺案”。这个案子还被媒体和学术界称为中央依法治国，保护产权的典型案例。可谁又知道，这只是一个被人为操纵炮制出来的噱

Figure 14: Letter allegedly written by Wang Linqing

The highest percentage of event-related censored images was related to a scandal involving China's Supreme People's Court. In late December 2018, prominent former television host Cui Yongyuan posted a series of posts on microblogging platform Weibo about the suspicious disappearance of several Supreme Court documents regarding a 2016 dispute over billions of dollars' worth of mining riches. According to Cui, the tips came from former Chinese Supreme Court judge Wang Linqing, who claimed that he was the first to find out about the missing court papers. Wang and Cui's revelation created controversy in which the Supreme Court first dismissed their words as "rumours" but soon said that it would investigate the case. It caught national attention not only because powerful figures from the highest echelons of the Chinese Communist Party were involved, including current Chief Justice Zhou Qiang, but also because of how the case evolved. In a shocking twist, whistleblower Wang Linqing said that he was the one who stole the papers. We found 24 images referencing the scandal, almost all of which are photocopies of a letter allegedly written by Wang Linqing, in which he accused Chief Justice Zhou Qiang of direct involvement in the disappearance of the court papers and illegal interference in the mining case in dispute (see Figure 14 for an example).



Global NEWS
National
Change Location
TV News Programs
Newscasts & Videos

World Canada Local Politics Smart Living Money Entertainment Health Video Podcast

CANADA December 11, 2018 12:47 pm

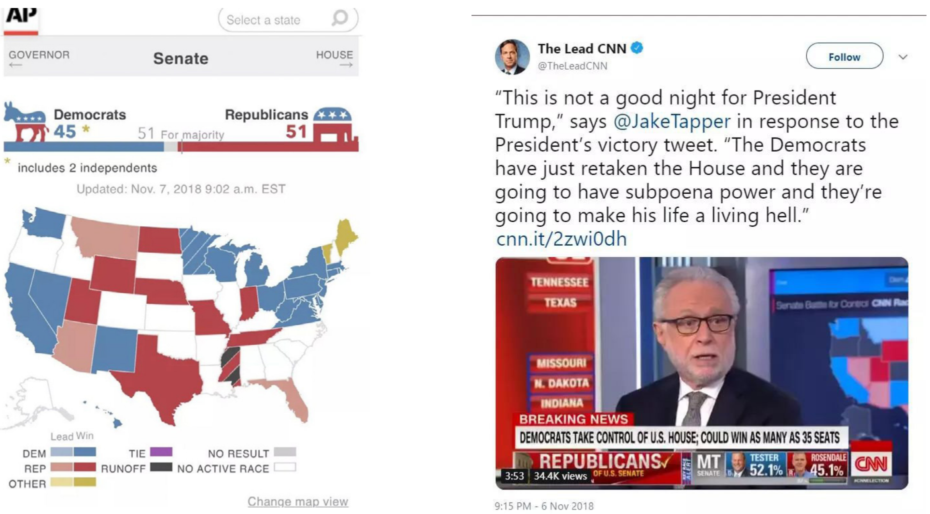
Arrest of Huawei's Meng Wanzhou is not political, U.S. ambassador to Canada says

By Staff The Canadian Press

Comments 12 | Facebook 81 | Twitter | Email | Print

Figure 15: Censored images referencing the arrest of Meng Wanzhou.

We found that nearly all images related to the Huawei saga were referencing the arrest of Meng Wanzhou, Huawei's CFO and daughter of Huawei's CEO Ren Zhengfei. Meng was arrested on December 1, 2018 by the Canada Border Services Agency under US charges of fraud and violating international sanctions. Meng's arrest quickly escalated to a diplomatic incident between China, Canada, and the United States. We found some censored images that contained negative sentiment towards Meng or Huawei, including photos of demonstrations led by overseas Chinese in support of extraditing Meng to the US but also neutral news coverage of the case (see Figure 15).



The Lead CNN
@TheLeadCNN
Follow

"This is not a good night for President Trump," says @JakeTapper in response to the President's victory tweet. "The Democrats have just retaken the House and they are going to have subpoena power and they're going to make his life a living hell."
cnn.it/2zwi0dh

9:15 PM - 6 Nov 2018

Figure 16: Two censored images related to the 2018 US midterm elections.

We found multiple images related to the 2018 US midterm elections (see Figure 16). In November 2018, elections were held in the United States with hundreds of congressional, state, and local seats in contest. Leaked directives show that Chinese regulators banned live-streaming and live updates of US elections likely for fear that it would encourage citizens to discuss the future possibility of open and free elections. Although elections are held in China, they differ from those in many democratic countries in two significant ways. First, under China's System of Multi-party Cooperation and Political Consultation, China is a de facto single-party state with other parties only having nominal representation. Second, elections in China feature limited direct election, with direct election being reserved to district-level or village-level positions. Elections of People's Congresses in China feature multiple levels including national, provincial, and often more than one municipal level, where each level except the bottommost is indirectly elected by the level below. The highest office, the President, is elected by the national-level People's Congress.

Social Issues

We found 32 images censored in the Social Issues category, which covers a range of content including the sale of illicit goods (e.g., see Figure 17), prurient interests, and non-political memes.

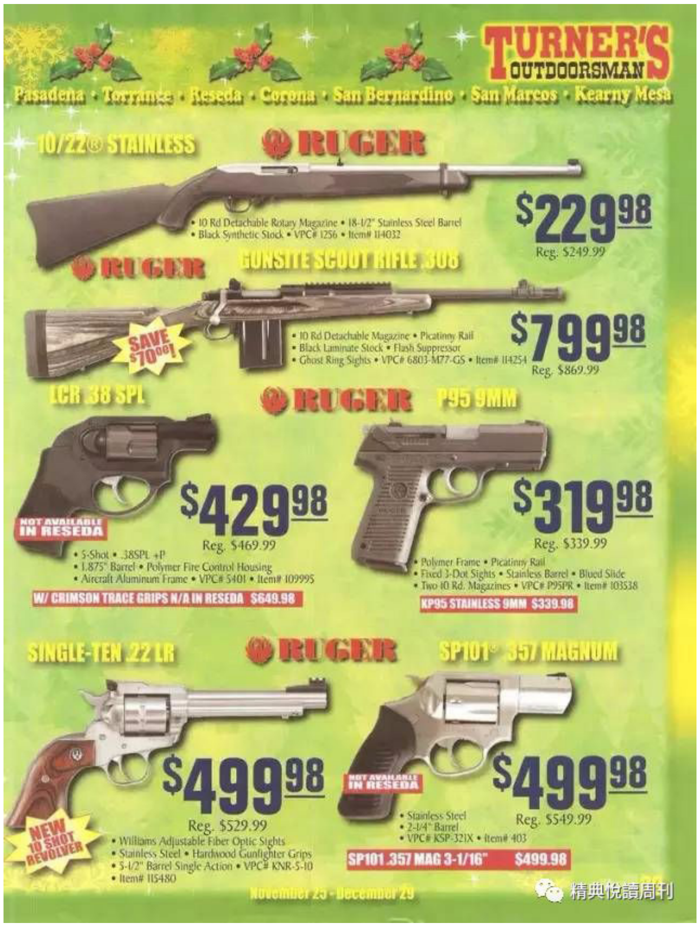

TURNER'S OUTDOORSMAN
Pasadena • Torrance • Reseda • Corona • San Bernardino • San Marcos • Kearny Mesa
10/22® STAINLESS
RUGER
• 10 Rd Detachable Rotary Magazine • 18-1/2" Stainless Steel Barrel
• Black Synthetic Stock • VPC# 1256 • Item# 114032
$229.98
Reg. $249.99
RUGER
GUNSITE SCOUT RIFLE .308
SAVE $70.00!
• 10 Rd Detachable Magazine • Picatinny Rail
• Black Laminate Stock • Flash Suppressor
• Ghost Ring Sights • VPC# 6803-M77-GS • Item# 114254
$799.98
Reg. $869.99
LCR 38 SPL
RUGER
P95 9MM
NOT AVAILABLE IN RESEDA
$429.98
Reg. $469.99
• 5-Shot • .38SPL +P
• 1.875" Barrel • Polymer Fire Control Housing
• Aircraft Aluminum Frame • VPC# 5401 • Item# 109995
W/ CRIMSON TRACE GRIPS N/A IN RESEDA $649.98
$319.98
Reg. $339.99
• Polymer Frame • Picatinny Rail
• Fixed 3-Dot Sights • Stainless Barrel • Blued Slide
• Two 10 Rd. Magazines • VPC# P95PR • Item# 103538
KP95 STAINLESS 9MM $339.98
SINGLE-TEN .22 LR
RUGER
SP101® 357 MAGNUM
$499.98
Reg. $529.99
NEW 10 SHOT REVOLVER
• Williams Adjustable Fiber Optic Sights
• Stainless Steel • Hardwood Gunfighter Grips
• 5-1/2" Barrel Single Action • VPC# KNR-5-10
• Item# 115480
NOT AVAILABLE IN RESEDA
$499.98
Reg. $549.99
• Stainless Steel
• 2-1/4" Barrel
• VPC# KSP-321X • Item# 403
SP101 .357 MAG 3-1/16" $499.98
November 25 - December 29
精典悅讀周刊

Figure 17: A flyer advertising firearms for sale (China does not allow the possession or sale of firearms).

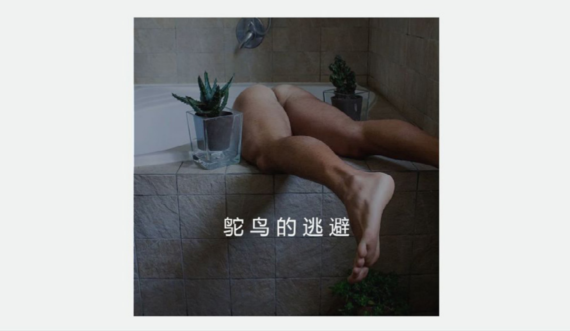


Figure 18: This image was seen in a WeChat article in which the author shared his experience of coming out to his mother; the text in the image describes "an ostrich's escape," which refers to the Chinese notion that ostriches bury their heads in the sand when facing troubles.

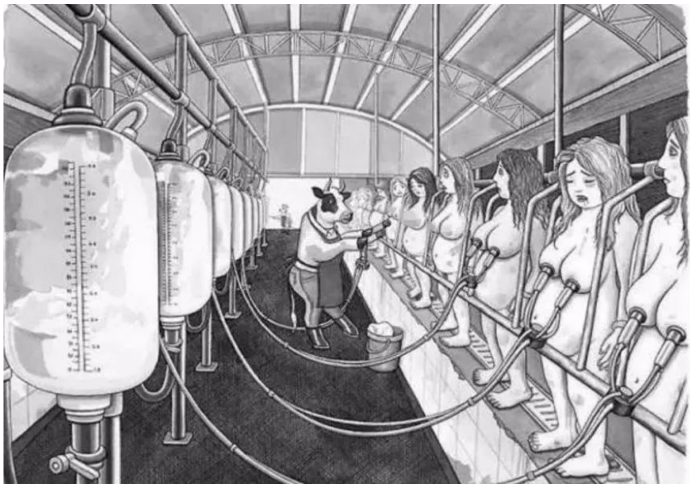


Figure 19: An image criticizing the exploitation of animals.

We found that much of the content involving nudity was likely blocked out of context. For example, the image in Figure 18 was extracted from a diary style article in which the author detailed his experience of coming out to his mother. The image in Figure 19 was from an article criticizing the exploitation of animals but was likely blocked due to sexually implicit content.

Social Resistance

Previous research suggests that Chinese social media platforms censor keywords pertaining to names of outspoken dissidents and advocates of social movement. We found 25 images in this category, including symbols and acts of resistance, from within China or overseas, against the Chinese government.

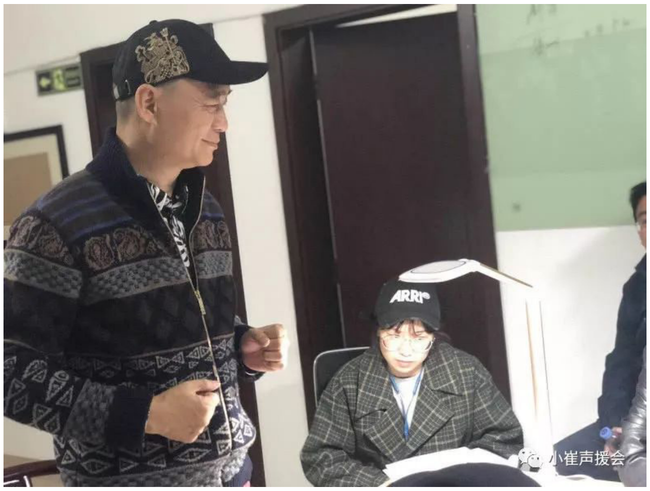


Figure 20: Photo of Cui Yongyuan, former television host and well-known social media commentator and whistle-blower.

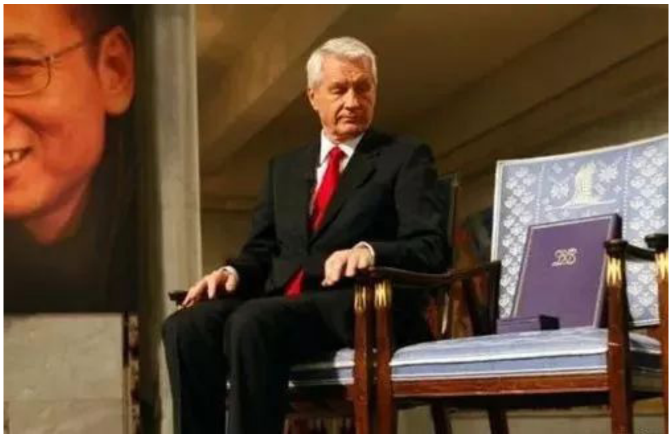


Figure 21: A chair reserved for Liu Xiaobo at the 2010 Nobel Peace Prize ceremony.

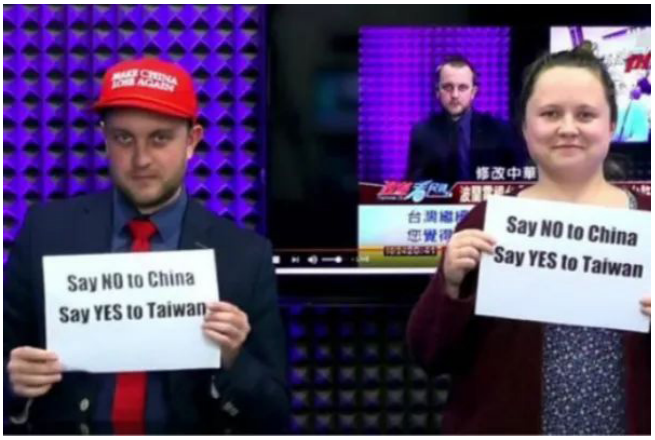


Figure 22: Hosts of a Polish YouTube channel express their support for Taiwan.

Among the 25 images, there are photos of individuals known for being outspoken about Chinese current affairs such as Cui Yongyuan (see Figure 20), screenshots of social media posts pertaining to online petitions, symbols of political resistance such as the iconic empty chair representing late Chinese Nobel Prize winner Liu Xiaobo (see Figure 21), as well as photos of overseas protests for Hong Kong and/or Taiwan against China's influence (see Figure 22).

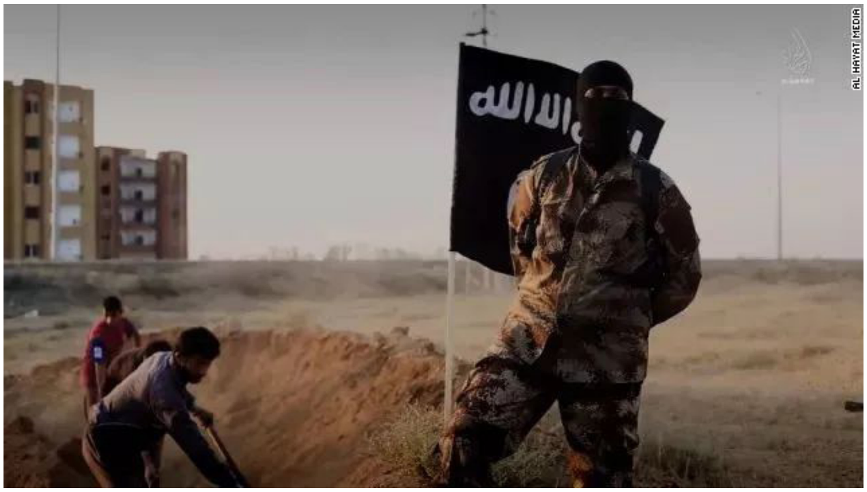


Figure 23: Image was extracted from a WeChat article criticizing the expansion of ISIS.

Other censored images include references to terrorism and religious extremism (See Figure 23 for an example) and financial phishing attacks.

Miscellaneous

Some images that we discovered censored had no clear reason for being on Tencent's blacklist. In our previous work studying keyword-based censorship, we found keywords with no clear reason for being blacklisted as well, and so we also expect a small number of blacklisted images to have no clear motivation for being blacklisted.

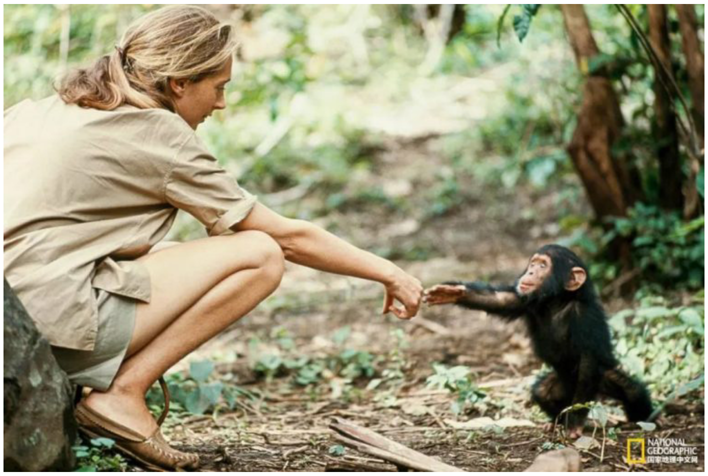


Figure 24: A photo of primatologist Jane Goodall and infant chimpanzee Flint.

One such case is a picture of famous primatologist Jane Goodall with an infant chimpanzee (see Figure 24), which we found blacklisted by Tencent. One possible reason for this image being sensitive is due to the use of chimpanzees in recent Chinese events as a racially derogatory reference, but the rationale for blacklisting this particular image is largely unclear.

Evaluating Qzone's versus WeChat's image blacklist

Measuring which images are filtered on Qzone provides insight into what Tencent broadly considers sensitive on their platforms but does not necessarily measure what is filtered on WeChat. To assess whether Qzone can be used as a proxy for testing image censorship on WeChat, we empirically tested whether Tencent's Qzone and WeChat platforms shared image blacklists.

To perform this evaluation, on April 22, 2019, we randomly sampled 30 images from the images we found filtered on Qzone and automatically re-uploaded them to Qzone to confirm that they were still filtered on Qzone on the date of testing. Among those still filtered, we manually uploaded them to WeChat Moments using one of our remaining test accounts. We then attempted to view each uploaded post from a different China-based account and measure which were deleted and which were not. We considered the two platforms consistent with respect to an image if and only if they both filtered an image.

We found that 29 out of 30 images (97%) that we had originally found filtered on Qzone were still filtered on Qzone at the time of testing. Among those 29, we found that 24 images (83%) that were censored on Qzone were also censored on WeChat Moments. These results show that image filtering on Qzone is largely effective at predicting filtering on WeChat Moments with few false positives.

Because we had previously found that some images filtered in WeChat group chat are not necessarily filtered in Moments, we decided to perform a second experiment comparing whether images filtered on Qzone were filtered on either Moments or group chat. In this experiment, we considered the Qzone and WeChat platforms consistent with respect to image filtering if an image found to be filtered on Qzone was also filtered on either WeChat Moments or group chat. While this is a broader definition, it takes into account that WeChat is not consistent with itself on many images.

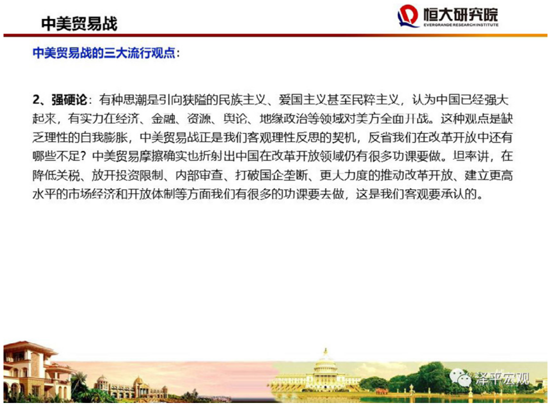
In this experiment, we found that 27 out of the 29 images filtered on Qzone (93%) were also filtered on either WeChat Moments or group chat. This finding shows that image filtering on Qzone is highly effective at predicting whether an image is censored somewhere on the WeChat platform with very few false positives.

It is not clear why some images are only filtered on Qzone as opposed to WeChat and vice versa. One explanation is that the blacklists are tailored according to which images are posted on each platform. Another explanation is that one platform is using an older or newer version of the same list as the other and that the two platforms may eventually become consistent with the same blacklist. A final explanation is that both platforms use the same image blacklist but that due to some sort of image processing done on one platform and not the other, some uploaded images after being processed no longer match the blacklisted image fingerprint on one platform versus the other.

Nevertheless, we found using Qzone as a proxy for measuring image filtering on WeChat to be effective. We suspect that this technique may be generalizable for future censorship research in any situation where two different platforms, possibly due to having the same operator, are suspected to use similar blacklists, and where one platform may be easier to access or measure than the other, despite possibly being not the original platform of interest.

Origin of blacklisted images

During our testing, we found that, for many censored WeChatscope articles, all images in the article were individually blacklisted, even if some of the images in the article were not themselves sensitive.



中美贸易战

恒大研究院
EVERGRANDE RESEARCH INSTITUTE

中美贸易战的三大流行观点：

2、强硬论：有种思潮是引向狭隘的民族主义、爱国主义甚至民粹主义，认为中国已经强大起来，有实力在经济、金融、资源、舆论、地缘政治等领域对美方全面开战。这种观点是缺乏理性的自我膨胀，中美贸易战正是我们客观理性反思的契机，反省我们在改革开放中还有哪些不足？中美贸易摩擦确实也折射出中国在改革开放领域仍有很多功课要做。坦率讲，在降低关税、放开投资限制、内部审查、打破国企垄断、更大力度的推动改革开放、建立更高水平的市场经济和开放体制等方面我们有很多的功课要去做，这是我们客观要承认的。

Figure 25: Screenshot of a presentation slide summarizing "popular views on Sino-US trade war"; the slides appear to be made by a research institute affiliated with the Evergrande Group, one of China's largest property developers.

In one instance, we found that five images were from the same WeChat public account article published by Global Times, a Chinese nationalistic tabloid newspaper owned by People's Daily. These images were not necessarily negative towards Huawei and the original article was criticizing overseas Chinese dissidents for being anti-China. In previous research, we found that Tencent applied broad censorship to sensitive events, restricting not only negative information about the government but also neutral references to government policy and screenshots of official announcements accessible via government websites. It is likely that Tencent over-censored images related to an event that it perceives as sensitive to the government or would cause unwanted public attention.

In another instance, we observed this same pattern in an article related to the so-called trade war between China and the US. In this article we found five images blacklisted, even though none were seemingly sensitive (see Figure 25 for an example).

These observations lead us to hypothesize that many of the images that Tencent blacklists are gathered from sensitive WeChat Public Post articles. If true, it is still unclear whether or not such images represent a large proportion of Tencent's image blacklist as a whole.

Conclusion

In this work, we study how Tencent implements image filtering on WeChat. We found that Tencent implements realtime, automatic censorship of chat images on WeChat based on what text is in an image and based on an image's visual similarity to those on a blacklist. Tencent facilitates realtime filtering by maintaining a hash index of MD5 hashes of sensitive image files.

We set out many of the properties of their hash index implementation, but we were most surprised to find that this hash database was populated by images that users, including both ordinary users as well as researchers, send over the platform. As researchers this leads to challenges wherein performing measurements can change the result of future measurements. It reminds us to be cautious and aware of that, even in systems of automated filtering, past measurements can change the future behaviour of filtering on the platform.

Our topic analysis of Tencent’s image content filtering identified 220 filtered images. These images were largely related to recent events and other political or government-related topics. Although we found references to a number of domestic Chinese scandals filtered, we also found images related to the 2018 US midterm elections. While we might expect Chinese censors to be sensitive to domestic criticism, this reminds us that even the reference to an outside alternate form of governance may also be sensitive.

Due to recent difficulty in obtaining and maintaining WeChat accounts, we analyzed the topic of images filtered on WeChat by proxy by measuring image filtering on Qzone, another Tencent product. Using sample testing, we found this technique largely effective. We hope that this technique of measuring content filtering on an easy to measure platform using a similar filtering implementation as a more popular but difficult to measure platform generalizes to future applications.