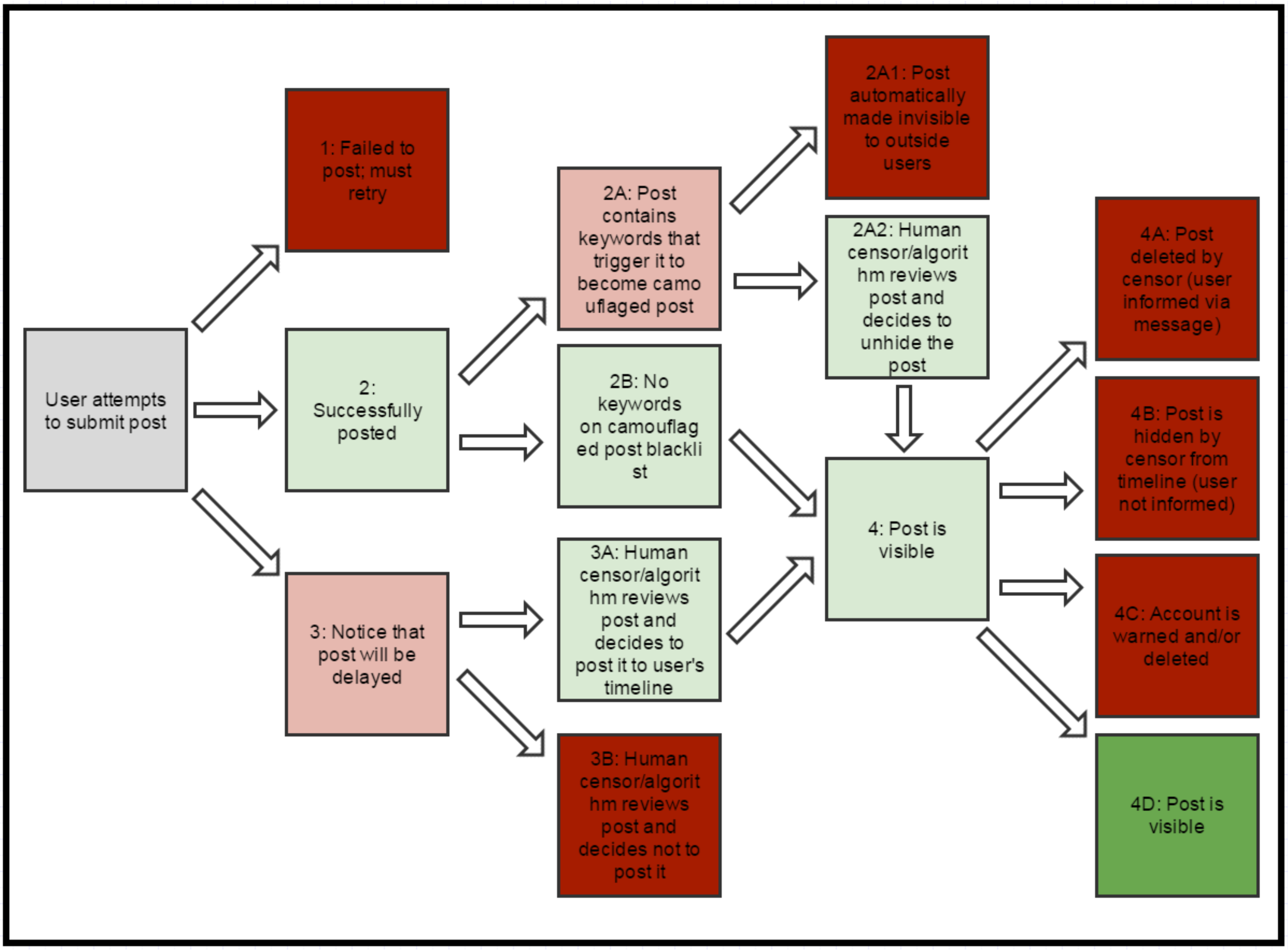
How a Weibo post gets censored

Jason Q Ng traces the path of a censored Weibo post and tracks keywords that trigger automatic review.


User attempts to submit post
1: Failed to post; must retry
2: Successfully posted
3: Notice that post will be delayed
2A: Post contains keywords that trigger it to become camo uflaged post
2B: No keywords on camouflag ed post blackli st
3A: Human censor/algorit hm reviews post and decides to post it to user's timeline
3B: Human censor/algorit hm reviews post and decides not to post it
2A1: Post automatically made invisible to outside users
2A2: Human censor/algorit hm reviews post and decides to unhide the post
4: Post is visible
4A: Post deleted by censor (user informed via message)
4B: Post is hidden by censor from timeline (user not informed)
4C: Account is warned and/or deleted
4D: Post is visible

Read about China's 2017 investigation into Weibo, WeChat and Baidu here

Part 1: Removal of censorship notice from Weibo searches

Internet censorship watchers have long noted the ubiquity of the phrase “???????????????????” on Chinese websites. That message (“According to relevant laws, regulations, and policies, search results cannot be displayed”) or one of its variants often accompanies search results that have been filtered or censored by Chinese search engines and content providers. While much of China’s censorship system is by design invisible or obfuscated, this notice has served both as a small beacon of transparency as well as a chilling reminder that authorities are in control of the information one sees on Chinese websites.

However, as of two weeks ago, Sina Weibo, a popular Chinese microblogging website, stopped explicitly reporting when its search results were censored, and instead, the site now returns a message stating that no results are found when a user searches for sensitive content. Search results for terms like ??? (Xi Jinping, the president of China) have long been filtered without the user's knowledge (returning only sanitized results) and users of the service oftentimes don't receive messages when their posts are rendered invisible (see Part 2). The removal of any acknowledgment of filtering during searches is another way of making censorship harder to easily detect.

Timing

This is not the first time Weibo has removed the censorship messages from searches for sensitive terms and reported "no results" instead. From November 1 to November 8 of 2012, Weibo used the same tactic to obscure its censorship. However, since November 2012 up until last month, Weibo has displayed the conventional censorship message, though with various increased efforts to filter its results during especially sensitive events like the June 4th anniversary in 2013.

According to our monitoring logs, this most recent removal began at roughly 9:05 pm EST on Oct 16 (10:05 am Beijing time on Oct 17). Since then, searches for sensitive keywords have returned zero results as opposed to a censorship message.

The timing of the first removal in 2012 occurred in the lead up to the 18th Party Congress, a once-a-decade transition of power in the Communist Party leadership and an especially sensitive event in China. Reports at the time noted of the extreme measures officials took to ensure the event would proceed smoothly—including the cancellation of theater performances, banning of kitchen knife sales, and the locking of taxi windows in Beijing. Following that logic, the decision to take extra measures to control the sharing and dissemination of criticism in the online sphere would make sense, and the removal of the censorship notice was likely only the most obvious change taken to prevent disruptions.

This latest removal again took place in the lead up to an important meeting of the Communist Party. The Fourth Plenum of the 18th CCP Central Committee occurred on October 20 to October 23, 2014. However, unlike the 2012 instance, censorship notices did not return to Weibo just before the start of the event, and instead, lasted through the duration of the meetings. It's possible the latest iteration may in fact be tied to another sensitive event: the APEC summit (Asia-Pacific Economic Cooperation) in Beijing on November 10 – 11, which world leaders like Barack Obama and Xi Jinping attended. Protesters currently occupying Hong Kong had initially suggested they might attend the summit, potentially making it even more sensitive, however, they have since postponed their visit to Beijing.

Analysis

To confirm the removal of the censorship message, we re-tested a sample of 2,429 sensitive keywords compiled by China Digital Times. This sample was last tested in its entirety on May 10 – 12, 2014, and yielded 822 keywords which were explicitly censored ("According to relevant laws, regulations, and policies, search results cannot be displayed"), 693 of whom had unique keywords which triggered their censorship (??, literally the numbers "six" and "four" refer to the crackdown on demonstrators in Beijing on June 4, 1989; ????, "June 4 Student Movement," is similarly censored, but is not considered a unique censored keyword). Sina Weibo indicated there were zero results for only ten of the tested keywords ("Sorry, unable to find related results").

We submitted the same 2,429 keywords through Weibo search engine on November 3 – 4. As expected based on preliminary testing, none of the keywords returned the censorship message. This time, there was a vast increase in keywords returning the "Sorry, unable to find related results" message (from 10 in our May test to 784). While a handful of those 784 keywords may genuinely have no results because they have not been used in Weibo posts, 754 of those 784 were in fact previously censored, a strong indication that searches for common terms yielding "no results" are in fact being censored.

As for the implications of the removal of the explicit censorship notice, we wrote previously in Tea Leaf Nation that:

Whether full-scale or piecemeal, the reduction of blanket keyword blocks is paradoxically a loss of transparency, since Chinese users no longer explicitly know when certain results are being specifically targeted for censorship . . . What is and is not off-limits has now become slightly harder to determine—another step in making censorship invisible and all-pervasive.

It will be worth following to see whether this change is permanent or merely temporary as in November 2012 and June 2013. If the explicit censorship notices return, we might surmise that tweaks to these censorship notices serve as a bellwether for particularly sensitive periods. More enhanced censorship may be occurring behind the scenes during these periods, but that will require more monitoring to confirm. Or it may simply be that Weibo is testing out new tactics.

Part 2: Pathways for censoring a Weibo post

A number of significant articles in recent years have done a great deal to illuminate how Weibo messages are censored before and after they are posted. Zhu, Phipps, Pridgen, Crandall, and Wallach's "The Velocity of Censorship: High-Fidelity Detection of Microblog Post Deletions" (presentation) noted that "Weibo has filtering mechanisms as a proactive, automated defense" and outlined these measures: explicit filtering (not allowing a user to post messages containing certain keywords), implicit filtering (alerting the user that a "data synchronization" issue would temporarily keep the message from being posted immediately), and camouflaged posts (a user is able to successfully post a message and it appears on their own timeline, but other users looking at that same timeline won't see it). These measures are not new; Jason Ng blogged about "vanishing posts" back in 2011 (English summary).

Bamman, O'Connor, and Smith ("Censorship and deletion practices in Chinese social media") along with King-wa Fu and his colleagues at the University of Hong Kong ("Assessing Censorship on Microblogs in China: Discriminatory Keyword Analysis and the Real-Name Registration Policy") have delved deeply via Weibo's API into identifying the deletion of posts after they've been submitted.

King, Pan, and Roberts recent Science article ("Reverse-engineering censorship in China: Randomized experimentation and participant observation") pays much needed attention to how employees working at Internet companies actually implement the systems for performing online censorship. The paper offers insight into the various options online content providers have for trying to ensure only acceptable material is viewable on the website, including the various "automated review" features.

As mentioned above, Weibo implements a number of automated review features, including many of those listed in Table 1 of King. Weibo posts also follow a similar lifespan to the pathways listed in Figure 1 of King, with some variations. Based on testing we performed (see Part 3), we identified the typical ways a message can be filtered or deleted on Weibo.

More explanation about some of the paths (see "Screenshots" below for images of these notices):

1) User submits a message with keywords on the explicit filtering blacklist. They receive a message "Sorry, this content violates … regulated regulations and policies. Operation cannot be specified." ("??????????????????(??)??????????????????????http://t.cn/8sYl7QG?") The user must remove the blacklisted keyword before being able to post. The submission is censored.

2A) If the post contains certain keywords, it may be automatically and instantaneously held for review by becoming invisible to all outside users. From the moment it is posted until it is reviewed by a censor, the post is invisible. From our two tests, it appears to take roughly 30 minutes before judgment is rendered on a post, but the amount of time no doubt varies depending on day of week and time of day.

3) User submits a message with keywords on the implicit filtering blacklist. They receive a message "Posted successfully. Please be patient about 1-2 minutes delay due to server synchronization, thank you." ("?????????????????????????????????1-2?????????")

3A/B) As noted by Zhu, this can sometimes take hours, but in other cases, can take about 30 minutes before the post—if approved (or not disapproved) by the censors—actually appears on a user's timeline.

4A/B) Quite rarely in our preliminary tests is a user actually informed that they have submitted unacceptable content and that it has been deleted (see Screenshot below). One of our accounts which did receive six of these notices received a warning notice and a 48 hour ban on posting (“???????????????????????????(??)?????????48????????http://t.cn/8s9ROKhc”), but no others did. Much more common is for posts to be rendered invisible (just as they were in 2A1) with no notice given to the user at all.

4C) Simply deleting a user’s account was quite common in our preliminary tests. It sometimes occurred hours after the account finished posting any messages, and sometimes occurred within seconds. Sometimes, the user was offered the opportunity to recover their account by submitting an appeal. However, due to the types of account used for the study and style of submitting content, we likely experienced much higher than typical rates of account deletion (as opposed to King who crafted genuine looking posts, thus likely evading banning due to certain filters looking for span and other non-standard user behavior).

Part 3: Keywords which trigger automatic review and project data

While both Zhu and King referred to these automated review mechanisms, they did not explicitly identify what specific keywords were on the blacklists that would trigger censorship via explicit filtering (box 1 in Figure 1), implicit filtering (box 3), and hiding of posts via “camouflage”/invisibility (box 2A). Below, are links to a Github repository with the data we have begun accumulating for tracking the keywords which trigger these three types of automatic review.

We performed this initial probe for the keywords which trigger these three methods of censorship by taking the 784 keywords which returned zero results from searching in the CDT list (words assumed to cause censored search results), and tried posting messages with them on Weibo. Unlike King, we did not craft realistic-looking messages. As we were just looking to identify words which would trigger automatic review, we simply posted the keyword along with a unique ID number for our own tracking purposes.

Because of rate limiting, we spread out the posting of these 784 keywords (which contained some duplicates) to 16 Weibo accounts registered to users in locations around the world, including Canada, the United States, France, England, Hong Kong, Macau, and mainland China. Our first test took place on November 8, Saturday morning, Beijing time and a follow-up test was run on November 10, Monday morning, Beijing time. (Each test was run using a computer in the United States; obviously, for future purposes, testing from different locations would be ideal to confirm there is no variation based on location.) We tracked the following:

- Whether a message with the keyword was not allowed to be posted (explicit filtering)
- Whether posting a keyword immediately returned a message stating that due to synchronization errors, the message would be posted later (implicit filtering)
- Whether a message with the keyword was immediately viewable by another user

(camouflaged post)

- Whether a message with the keyword was viewable by another user one hour after being posted and 24 hours being posted
- Whether or not a user received a message in their inbox from the administrators stating a post was being deleted
- ¥ Whether or not a user’s account was suspended

Caveat: the majority of the accounts we used to test for automatic review censorship were basic accounts at level 0. It’s possible another user with different privileges might have higher or lower degrees of censorship. However, we were able to consistently replicate the majority of the below results when using new accounts, and hopefully others can as well.

In our first test, 66 of the keywords could not be posted, including ??? (Xi Mingze, the daughter of Xi Jinping), numerous variants on Zhou Yongkang’s name (a former government official under investigation for corruption), and ??? (a derogatory play on the term for the 18th Communist Party Congress). In our second test, the number increased slightly to 73. Though it was not identical across the two tests, there was a large overlap, with 63 of the initial 66 also being unable to be posted on the second test.

In our first test, 14 of the keywords triggered implicit filtering, including ???? (Beijing authorities), ??? (Ding Zilin, an activist), and ???? (Chinese Spring, a play on Arab Spring). In our second test, the number increased slightly to 15. Again, though not identical, there was a large overlap, with all 14 also triggering the implicit filtering during the second test.

In our first test, 133 of the posts were not immediately viewable by another user (camouflaged posts). These included posts containing the keywords 5?35? (May 35, code for June 4), ??? (“hairy bacon,” an insult for Mao Zedong), ????? (Boxun News), and ??? (Tank Man). 202 were hidden/deleted within 1 hour, and 273 were hidden/deleted within 24 hours. Eighteen messages regarding deletions were received and four of the sixteen accounts used for the test ended up being suspended. The numbers were roughly the same during the second test. We are less confident of the actual total number of keywords which trigger posts being deleted/hidden after one hour and 24 hours in the second test because twice as many accounts were suspended during the test, including some before the first hour had passed. Thus a greater proportion of deleted/hidden posts were attributed to account suspension and unable to be confirmed as being due to the sensitivity of the keyword itself. That said, there was still a large overlap between the two tests with regards to camouflaged posts: only 19 of the 133 camouflaged posts from the first test were subsequently viewable on the second test. Further testing should be done to identify more precisely which keywords are causing posts to be automatically held for review in this manner.

As a rudimentary control, we also posted 16 keywords from the CDT sensitive words list that returned results (and thus, were not fully censored by Weibo search) in the November 3rd test. We also posted the 33 chapter titles (in Chinese) of Quotations from Chairman Mao and the opening of the Gettysburg Address (in English). Not one of these keywords triggered censorship either via automatic review or were hidden/deleted in the subsequent 24 hours in either test.

Jason Q Ng is a Research Fellow at The Citizen Lab. This report was originally published for The Citizen Lab on 10 November 2014. You can follow him on Twitter on @jasonqng.

Published on:December 5, 2014