Journal Title
XX(X):1–16
©The Author(s) 2016
Reprints and permission:
sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/ToBeAssigned
www.sagepub.com/
SAGE

Censorship's Effect on Incidental Exposure to Information: Evidence from Wikipedia *

Jennifer Pan[1] and Margaret E. Roberts[2]

Abstract

The fast growing body of research on internet censorship has examined the effects of censoring selective pieces of political information and the unintended consequences of censorship of entertainment. However, we know very little about the broader consequences of coarse censorship, or censorship that affects a large array of information like an entire website or search engine. In this study, we use China's complete block of Chinese language Wikipedia (zh.wikipedia.org) on May 19, 2015 to disaggregate the effects of coarse censorship on proactive consumption of information – information users seek out – and on incidental consumption of information – information users are not actively seeking but consume when they happen to come across it. We quantify the effects of censorship of Wikipedia not only on proactive information consumption, but also on opportunities for exploration and incidental consumption of information. We find that users from mainland China were much more likely to consume information on Wikipedia about politics and history incidentally rather than proactively, suggesting that the effects of censorship on incidental information access may be politically significant.

Keywords: censorship, Wikipedia, China, information consumption

Introduction

Wikipedia—a wiki-based website where users collaboratively modify content and structure—is one of the most widely viewed sites in the world.[1] As of November 2018, Wikipedia had over 49 million pages in nearly 300 languages.[2] Although Wikipedia content is user generated and changes over time, studies have consistently shown Wikipedia to be an accurate source of a very wide variety of information.[3]

In this paper, we use China's complete block of Chinese language Wikipedia (zh.wikipedia.org) on May 19, 2015 to understand how coarse censorship—censorship that is not selectively aimed at suppressing one specific type of content—influences the overall consumption of information. We distinguish between two ways in which people consume information—proactive consumption, when users know what information they want and actively seek it out, and incidental consumption, when users encounter and consume information they were not proactively seeking. We find that censorship not only affects information users seek out proactively but also has dramatic effects on incidental consumption of information – in this case, information that Wikipedia users accessed through the homepage. Further, we show that Wikipedia users from mainland China were more likely to encounter political and historical information incidentally rather than proactively. These results imply that coarse censorship can have long-range consequences by cutting off opportunities for exploration and by-chance encounters with information, and can suppress consumption of political information that people may not know they demand. These results join an emerging strand of research on the broader consequences of censorship (1; 2) that augments studies of the effects of selective censorship of political information (3; 4; 5; 6; 7; 8; 9; 10) and the unintended consequences of entertainment-related censorship (11; 12)

[1] Assistant Professor, Department of Communication, Stanford University, jp1@stanford.edu
[2] Associate Professor, Department of Political Science, University of California San Diego, meroberts@ucsd.edu

We construct a timeline of Wikipedia page views for each of 372,208 pages that allows us to approximate the total number of page views for each page accessed through proactive consumption and the number accessed through incidental consumption through the Wikipedia homepage. While our data do not allow us to differentiate directly between mainland and non-mainland consumption of information, we use the sudden nature of the Wikipedia block to estimate the number of proactive and incidental page views originating from mainland China as opposed to other locations around the world. Using automated methods to analyze the content of these Chinese Wikipedia pages, we find that exploration on Wikipedia via the homepage inspired mainland users to incidentally consume a broad array of information, particularly about the cultures, histories, and politics of countries beyond China. In contrast, proactive consumption of information on Chinese Wikipedia brought mainland users largely to entertainment and scientific pages.

The next section describes the background of China's block of Wikipedia and the impact of the block on overall page views. In the third section, we describe the impact of the homepage on page views. Then, we describe how we decompose views of an individual page into proactive views and views that were generated incidentally by the homepage as well as differentiate between mainland and non-mainland viewers. The fifth section uses a topic model to describe the types of pages that were viewed by mainland users incidentally versus proactively, and the last section concludes.

Overall Impact of Wikipedia Censorship in China

Chinese language Wikipedia was launched in 2001. Although media in China is tightly controlled (13; 14), Chinese language Wikipedia pages were largely available in mainland China until May 2015. From 2004 to 2008, China occasionally blocked all access to Chinese language Wikipedia with the Great Firewall.[4] During these blackouts, which ranged from a few days to a few months, it was not possible to access Wikipedia from a mainland China IP addresses. After 2008, Wikipedia pages with politically sensitive content—such as pages related to the 1989 Tiananmen Square protests, political activists, or controversial historical events—were selectively blocked. The majority of Chinese language Wikipedia pages remained accessible to users from mainland China, including the preponderance of political and historical information.[5]

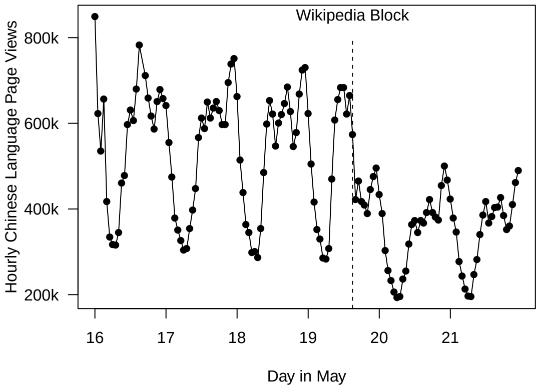

Figure 1. Page views of Chinese language Wikipedia by hour, May 16-21, 2015. The Wikipedia block occurred during the afternoon of May 19, 2015.

In 2011, Wikipedia added support for Hyper Text Transfer Protocol Secure (HTTPS), which allows data to be transferred with encryption.[6] HTTPS prevents internet service providers (ISPs) and governments that control ISPs from seeing what specific page users are visiting in any particular web domain. This prevented the Chinese government from selectively blocking particular Wikipedia pages for users who were using the HTTPS version. In response, China began blocking access to the HTTPS version of Chinese Wikipedia in 2013 while allowing the un-encrypted HTTP version to remain available. Because the HTTP version was not encrypted, the government could see what pages were being visited, and continue to selectively block the pages they deemed objectionable.

In 2015, Wikipedia made encryption mandatory by redirecting all HTTP requests to the corresponding HTTPS addresses. As a result, Chinese authorities could no longer determine what pages users were viewing and could no longer selectively block pages.[7] On May 19, 2015, China began a wholesale block of Chinese language Wikipedia that continues as of the end of 2018.

News reports mention only that the block occurred on May 19, 2015, but hourly page view data made available by the Wikimedia Analytics team[8] clearly shows that the block occurred sometime within the hour of 3:00PM China time on May 19. Figure 1 shows views of Chinese Wikipedia by hour from May 16 through May 21. Between May 16 and May 18, we see the usual daily pattern in Chinese language Wikipedia views, which drops during the early hours of the morning (between 1AM and 5AM China time) and then rises during the day and evening. This pattern is interrupted mid-day—between 3PM and 4PM China time—on May 19, when page views experience a sudden decrease and continue at a much lower rate through the end of May.[9]

The page views of Chinese language Wikipedia pages that remain after the block represent page views generated by Chinese speakers in areas not affected by China's Great Firewall, such as Hong Kong and Taiwan.[10] While some page views of Chinese language Wikipedia after the block may originate from mainland Chinese users who "jump" the Great Firewall by using a Virtual Private Network, this number is likely small for two reasons. First, there are few VPN users in mainland China (2; 1), and second, the Wikipedia block did not (unlike China's block of Instagram) lead to an increase in VPN downloads or installations in mainland China (2).

The Wikipedia block had an enormous impact on the number of page views, indicating that the block significantly affected the number of people from mainland China reading material on Chinese language Wikipedia. There were 664,694 page views of Chinese language Wikipedia pages between 2PM and 3PM on May 19th, directly before the block. In comparison, there were 421,663 total page views between 4PM and 5PM on May 19th, resulting in decrease of 243,031 page views between these two hourly segments. These numbers suggest that around 35% of the Chinese language Wikipedia page views before the block originated from mainland China. The block decreased views of Chinese language Wikipedia originating from mainland China on the order of 3 million page views per day. If this same trend had continued through the end of 2018, that would mean it has decreased page views cumulatively on the order of 3 billion page views.

Homepage Effects on Page Views

We can use the sudden nature of the block of Wikipedia from mainland China to better understand the content that mainland China users were viewing on Wikipedia and how they arrived at this content. Because the full block of Wikipedia was motivated by a technological change to the platform of Wikipedia rather than by any particular event, we expect that users were using Chinese language Wikipedia as they normally would on the day of the block. Thus, the full block of Wikipedia on May 19, 2015 provides us with a unique opportunity to estimate how mainland Chinese users were consuming information, and, by extension, evaluate the impact of this coarse censorship on information consumption.

To do this, we focused our analysis on a subset of 372,208 pages that were viewed on at least two days in the 2PM-3PM hour the week prior to the block. We do this to exclude many pages in the Wikipedia dataset that are blank pages, which are likely to be viewed only once in the week before the block.[11] For a preliminary, cursory idea of the pages that mainland viewers were looking at most on May 19, 2015, we generate a list of the pages most affected in terms of total page views by the block. We computed the per page difference between page views in the hour before the block (2-3PM on May 19) and after the block (4-5PM on May 19) and selected the 100 pages with the largest page view decreases. While the top pages only provide a partial picture of the Wikipedia content affected by the block, the 100 pages (0.018% of all pages) that lost the most views account for over 10% of the total page view loss, or a loss of almost 30,000 views over the hour. A table of these pages is provided in the Appendix (Figure 4).

At first glance, many of the pages most affected by the block are surprisingly unrelated to ongoing events and not facts that we would have expected to be of widespread interest during this time period to anyone in mainland China. For example, the page for "Susan B Anthony" lost 139 views between the hours of 3PM and 5PM, and was the 47th most affected page in the dataset. "Ghost marriage" a traditional custom lost 196 views, the 25th most affected page in the dataset. However, the common feature of many of the pages most affected by the block is that they were all pages shown to users through the Chinese Wikipedia homepage on May 19, 2015. The Wikipedia homepage features a collection of highlighted articles, related to events that happened on the same date in history, random facts in a 'did you know?' section, people and places that are in the news, and sets of images.[12]

In total, 65 out of the 100 most affected pages by the block were linked from the homepage – 11 pages were linked directly to the homepage, 26 were linked to links from the homepage, and 28 were links of links of links from the homepage, suggesting that exploration of Wikipedia content through the homepage was an important source of page views from mainland China and that censorship had an important impact on incidental consumption of information in mainland China.

In order to systematically estimate the impact of the homepage on mainland views on Wikipedia, we document all pages in our dataset that were linked to the homepage on May 19th. Unfortunately, we cannot recover the complete content of Wikipedia homepages and other browsing related pages because their main content is not user generated (user edits only apply to the formatting of these pages). However, some of the sections of the home page are archived in other Wikipedia pages, allowing us to recreate almost all of the May 19 homepage of Chinese Wikipedia. Specifically, Chinese Wikipedia has a "Good Entries" archive (Wikipedia:優良條目/2015年5月) that lists featured articles by day from March 1, 2012 onward, as well as a "Special Entries" archives (Wikipedia:特色條目/2015年5月) that does something similar. Wikipedia maintains a list of events that happened "on this day" in history that are then used on the homepage: Wikipedia:史上的今天/5月 as well as an archive of facts of newly created Wikipedia pages from the "Did you know section": Wikipedia:新目推荐/2015年5月.

In addition to identifying links of articles from the Wikipedia homepage, we collect secondary and tertiary links to the homepage. Given that the Wikipedia homepage was associated with large declines in page views, we should expect that links to the links included in the Wikipedia homepage and links of the links of the links in the Wikipedia homepage may also have been disproportionately affected by censorship. If mainland users were finding content to consume on Wikipedia through the homepage, then they may follow links within those linked pages. Therefore, in addition to documenting pages that were linked to the homepage, we use the Wikipedia API to identify all links of the links of the homepage and links of the links of the links of the homepage.

We then compute the average loss in page views between 2-3PM and 4-5PM for those pages linked from the homepage, those linked to pages linked from the homepage, and those linked to links of links from the homepage.[13] We report these statistics in Table 1. We find that the average loss in page views is largest for pages linked to the homepage, second largest for those linked to links on the homepage, and third largest for those linked to links. For pages without a link from the homepage, the average decrease in page views is smallest. Overall, pages that are directly or indirectly linked to the homepage account for 54% of the decrease in page views, suggesting that a significant portion of page views of Wikipedia originating from mainland China could have been driven by the homepage.

	Average Decrease	Number of Pages
Links from homepage	-45.07	45
Links of links	-2.06	7,388
Links of links of links	-0.58	115,353
No links from homepage	-0.29	249,422

Table 1. Average per page decrease in page views before and after block, for pages linked to the homepage, pages linked to links from the homepage, pages linked to links of links from the homepage, and pages with no links from the homepage. Pages that experienced the largest average decrease in page views on average were those with links to the homepage.

Could These Effects Be Due to Crawlers?

In using page view data, one may be concerned that page views are generated by random crawlers rather than real individuals. Specifically, crawlers may be programmed to start at the Chinese Wikipedia homepage each day and then go to each linked page, which could explain the patterns we documented above.

The Wikimedia Foundation provide information on the likely type of agent—user, spider, bot—accessing each page over time only beginning in July 2015.[14] For the month of July 2015, 1.7% of all views of the Chinese Wikipedia was made by bots. If we look at the page views of homepages of all wikipedia projects with more than 1 million pages, on average 0.16% of all page views were made by bots in July 2015. While this data is from after the block, this suggests bots and crawlers are not driving our results.

Two other pieces of evidence suggest that these page view patterns are not driven by bots. First, if bots and crawlers dominated, we would expect all links on the homepage or all links in a section of the homepage to decrease at similar rates before and after the block. However, decreases in page views to pages linked to the home page varied, and the pages with the largest decrease in page views were not always those at the top of the home page. This suggests that even if crawlers account for some of the loss of page views after the block, they do not describe the entire picture. Second, while there might be some crawlers on Chinese Wikipedia from China, most crawlers would likely be from IP addresses outside of the United States, which are unaffected by the Great Firewall. Even if a crawler originated from inside mainland China, the crawler would likely access Chinese Wikipedia through a VPN in order not to be selectively limited by the Great Firewall (prior to the wholesale block). Wikipedia crawlers from outside of China would not have been affected by the May 19 block. Therefore, we are less concerned that our results are affected by crawlers than we would be if we were simply measuring overall views of pages rather than their relative decrease.

Decomposing Incidental and Proactive Views By Page

We can not know for certain that all of the page views of the pages directly or indirectly linked to the homepage were in fact viewed as a result of the homepage. It could be that some of these pages were popular in mainland China independent of their appearance of the homepage. In this section, we use the time series of each page linked to the homepage to estimate the extent to which the decrease in page views for each page was driven by its appearance on the homepage—which we call *incidental* page views—versus the page's popularity outside of it's appearance on the home page—which we will call *proactive* page views.

We use the example of the Chinese language page for the U.S. one dollar coin (1美元硬) as an illustrative example of how we can estimate which views were generated by the homepage and which views reflect the popularity of the page outside of the homepage. Overall, the U.S. one dollar coin was not a very popular page on May 18th, with total page views hovering around 10 per hour the day before it was featured on the homepage. However on May 19th, the page was featured on the homepage and was viewed a whole lot more—150 times in the 2-3PM hour before the block. Assuming that nothing else changed except its placement on the homepage, the difference between the hourly page views per page i on May 19th in the 2-3PM hour ($V_{i,2pm,5/19}$) and May 18th in the 2-3PM hour ($V_{i,2pm,5/18}$) is an estimate of the total increase in page views for the U.S. one dollar coin page because of the homepage—page views driven by incidental consumption—for the 2-3PM hour ($V_{incidental}$), as shown in Equation 0.1.[†]

$$V_{i,incidental} = max(V_{i,2pm,5/19} \tag{0.1}$$

$$-V_{i,2pm,5/18}, 0) \tag{0.2}$$

Because we cannot distinguish before the block between mainland and non-mainland viewers, this difference, $V_{i,incidental}$, includes incidental consumption from both mainland China users (m) and non-mainland China users (m'). Sometime during the 3-4PM hour, Chinese Wikipedia page views from mainland China were blocked, leading to a large decrease in page views of the U.S. one dollar coin page between the hours of 2-3PM and 4-5PM, when page views dipped to around 25 views. Page views further decrease on May 20th to around 5 views per hour, when Chinese language Wikipedia is still blocked, and the U.S. one dollar coin is no longer featured on the Wikipedia homepage. Because the block only affected mainland Chinese users, the difference between the 4-5PM page views on May 19th ($V_{i,4pm,5/19}$) and 4-5PM page views on May 20th ($V_{i,4pm,5/20}$), is an estimate of the incidental consumption generated by the homepage for non-mainland viewers only ($V_{i,incidental,m'}$). Non-mainland incidental consumption of pages linked directly or indirectly to the homepage can then be estimated as:

$$V_{i,incidental,m'} = max(V_{i,4pm,5/19} \tag{0.3}$$

$$-V_{i,4pm,5/20}, 0) \tag{0.4}$$

Using the estimates from Equation 0.1 and Equation 0.3, we can then obtain an estimate for incidental consumption from mainland China. Incidental consumption from mainland China is the difference between $V_{i,incidental}$ and $V_{i,incidental,m'}$ for each page linked directly or indirectly to the homepage.

$$V_{i,incidental,m} = max(V_{i,incidental} \tag{0.5}$$

$$-V_{i,incidental,m'}, 0) \tag{0.6}$$

Total mainland consumption is the difference in page views in the hour before ($V_{i,2pm,5/19}$) and hour after ($V_{i,4pm,5/19}$) the block:

$$V_{i,total,m} = max(V_{i,2pm,5/19} \tag{0.7}$$

$$-V_{i,4pm,5/19}, 0) \tag{0.8}$$

To ensure that we complete the decomposition, we do not allow the estimated number of incidental consumption from mainland China to exceed our estimate of total mainland consumption. We impose the restriction:

$$V_{i,incidental,m} = min(V_{i,total,m}, \tag{0.9}$$

$$V_{i,incidental,m}) \tag{0.10}$$

Finally, we can estimate total proactive consumption from mainland China ($V_{i,proactive,m}$) by subtracting incidental mainland consumption ($V_{i,incidental,m}$) from total mainland consumption ($V_{i,total,m}$).

$$V_{i,proactive,m} = V_{i,total,m} - V_{i,incidental,m} \tag{0.11}$$

Altogether, we use this calculation to roughly estimate the total page views of each page coming

[†] Since views cannot be negative, we take the maximum of this difference and zero.

from mainland China, as well as decompose views of each page into page views generated by the homepage (incidental consumption) versus page views that were accessed through other means (proactive consumption). We acknowledge that what we count as proactive consumption may also include incidental consumption, through means outside the homepage. For example, it is possible for someone to arrive at a Wikipedia page because they were reading a blog post that linked to that page instead of proactive searching or seeking that information. Since our results are focused on incidental consumption, this bias in the data is more likely to work against us.[15]

From this decomposition, we estimate that that approximately 25% of the loss in page views due to the block were page views based on incidental consumption – views it received the day it was linked to the homepage above and beyond what the page generally received. This indicates that while censorship did have an impact on limiting information users were proactively looking for, its impact on incidental consumption of information was also very important, accounting for a significant number of total page views.

Topical Differences in Proactive vs. Incidental Consumption

We now turn to the question of the types of content mainland viewers were seeking out proactively, in comparison to the types of content that mainland viewers were consuming incidentally because of what was featured on the homepage. We find that while on average mainland viewers proactively sought out information about entertainment and scientific facts, the Wikipedia homepage facilitated incidental page views of political and historical information as well as information about other countries and cultures.

In order to identify the topics that mainland users were proactively consuming versus those they were incidentally consuming, we use a topic model to describe the topics of the pages that were being viewed before and after the block and to estimate which topics were most associated with incidental versus proactive page views. The idea behind statistical topic models is that they can inductively identify clusters of words, or “topics," within the text that are commonly used together (15; 16; 17; 18). Each document is then made up of a combination of topics. Thus the two main outputs of the model are the words likely to appear in each topic (topical content), and the amount each topic appears in each document (topical prevalence).

Using a topic model to describe the content of the pages that were viewed before and after the block allows us to group the pages into topics or themes that were most affected by the block.[16] Using the entire text of each Wikipedia page as an input to the topic model would give us a very different amount of text per page. Instead, we used the Wikipedia API to extract the summary information for the pages in our dataset. The summary information for each Wikipedia page is the first paragraph of the page and exists for the majority of pages. The summary information is ideal for our setting because it is relatively consistent in length across pages and typically contains the key information about the content of the page.

We queried the API for all pages within our dataset that were viewed in the hour directly before or after the May 19th block.[17] We excluded any pages that did not have any content or were too short to contain summary information. We also excluded pages that were not viewed in the hour before or hour after the block. In total, our analysis contains 158,611 summaries. We used the Structural Topic Model (STM) to describe the topics of these summaries (19). We used an automated method to select the number of topics (20). In total, our model identified 78 total topics across the 158,611 pages, which we labeled after reading the highest probability words within the topic and example documents from each topic.

The topic model describes the kinds of pages that mainland Chinese readers of Wikipedia were viewing before versus after the block. Because the list of topics is extensive, we refer the reader to the Appendix for full details on the topics (Table 2), where we report the highest probability words, estimated total page views originating from mainland China per hour, and the proportion of the page views for each topic from browsing. The 78 topics are mostly organized around Substantive and Entertainment related content, outside of two general words topics and one topic about Wikipedia Entries. Sixteen topics are mainly related to entertainment, including television, celebrities, music, sports, novels, video games, and pornography. The remaining 50 topics are substantive topics unrelated to entertainment—for example, European history, computer systems, Qing and Ming dynasty history, and business and the stock market. In the Appendix, we also report the estimated total page views per hour from mainland China ($V_{i,total,m}$). Outside of general word topics, the most popular topics in mainland China include many topics about history and politics, including Democracy, Region Statistics, and the Government and Legal System in China.

But were the most popular topics viewed by mainland users of Wikipedia sought out proactively or happened upon incidentally? For each page $i = 1, ...N$, the topic model outputs the proportion of the summary in each of the 78 topics, $\vec{\theta_i}$. To measure on average how many page views in each topic that were generated from incidental consumption, we multiply the topic proportions for each page $\vec{\theta_i}$ by the estimated incidental page views from the mainland for that page $V_{i,incidental,m}$. We do the same for proactive page views, multiplying the topic proportions for each page $\vec{\theta_i}$ by the estimated proactive page views from the mainland for that page $V_{i,proactive,m}$. We then take the mean of all proactive and incidental page views weighted by topic to estimate the proportion of incidental page views in each topic:

$$\vec{T}_{incidental} = \frac{\sum_{i=1}^{N} \vec{\theta_i} * V_{i,incidental,m}}{\sum_{i=1}^{N} V_{i,incidental,m}}$$

and the proportion of proactive page views in each topic:

$$\vec{T}_{proactive} = \frac{\sum_{i=1}^{N} \vec{\theta_i} * V_{i,proactive,m}}{\sum_{i=1}^{N} V_{i,proactive,m}}$$

We take the difference $\vec{T}_{incidental} - \vec{T}_{proactive}$ to describe which topics are disproportionately viewed incidentally versus proactively.[18] Figure 2 shows the topics most associated with proactive page views and Figure 3 shows the topics most associated with incidental page views. Overall, we find that proactive page views were disproportionately driven by entertainment, including information about Korean pop music, animation, Japanese celebrities, music, and sports. Incidental page views, on the other hand, were disproportionately focused on pages that talked about politics and history, ranging from the Chinese Communist Party, and government and legal system of China, o the histories of countries around the world to the politics and geography of other countries.

Our analysis suggests that views of pages discussing political and historical topics in mainland China was driven largely by incidental consumption, likely facilitated by the fact that the Wikipedia homepage includes many links to political and historical topics. This evidence suggests that users became interested in political and historical information when they happened upon it on Wikipedia, even when they did not seek this information out directly.

This has several important implications. First, political information might be disproportionately affected by censorship that affects exploration and happenstance encounters with information, and coarse censorship of a broad website like Wikipedia has political implications above and beyond what users actively seek out on the Internet. Much of the demand for political information may be endogenous, facilitated by the platform, rather than sought out by the user. This points to the importance of general exploratory platforms that point users to new topics.

Second, this could mean that users could do little to compensate for the political and historical content censored by the Wikipedia block. It is straight forward for users to seek out information that they queried proactively on Wikipedia on other websites. However, users would not know what information that they would have arrived at on Chinese language Wikipedia incidentally, and this information would be very difficult to seek out out other websites without Wikipedia, since they would not know the pages that they would come across if they had access to the full Wikipedia website. Even when there are substitutes (in this case the website Baidu Baike in China), the value added of sites with exploratory pages goes above and beyond specific information users seek out.

Conclusion

In this paper, we use China's complete block of Chinese language Wikipedia on May 19, 2015 to examine the impact of coarse censorship. We disaggregate proactive consumption from incidental consumption of information on Wikipedia, and we find that coarse censorship affects consumption of information by cutting off opportunities for exploration and incidental consumption of information. This decrease is important because users were much more likely to consume information about politics and history incidentally rather than proactively. Proactive consumption of information focused on entertainment topics, as well as scientific topics. We speculate that the proactive consumption of scientific information may be related to either students needing specific information for assignments and projects, or working professional needs specific information for their jobs.

These results have implications for how censorship can limit the creation of an informed and critical public. People do not know what they do not know, and cannot demand or search for information they are unaware of. Coarse censorship impeding access to platforms that facilitate incidental consumption of information limits the public's ability to encounter previously unknown information. Given Internet users' impatience on the web and unwillingness to

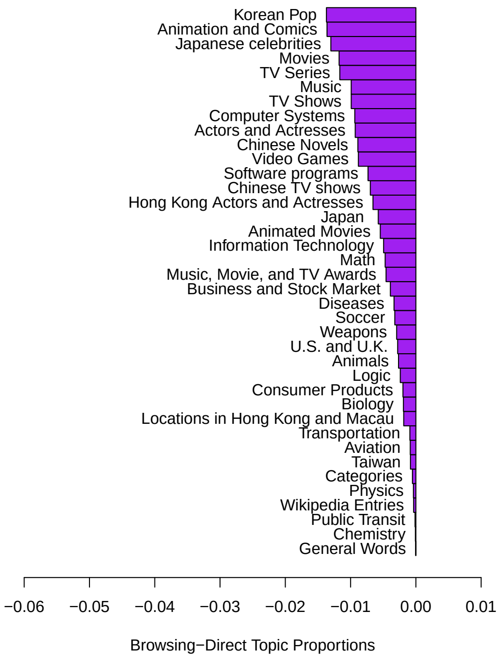

Figure 2. Topics associated with proactive page views, estimated from the Structural Topic Model. The length of the bar shows the difference in topic proportions between incidental page views and proactive page views.

overcome the costs of censorship to seek out information (2; 1), greater impediments to exploration may endogenously decrease proactive demand for certain types of information because users are less likely to come across a greater variety of information incidentally.

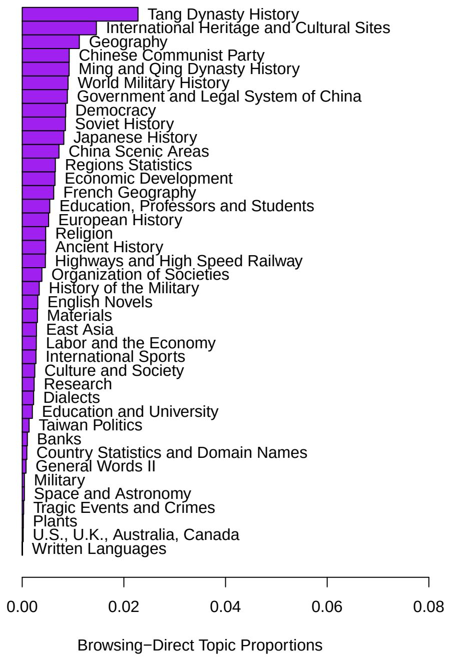

Figure 3. Topics associated with incidental page views, estimated from the Structural Topic Model. The length of the bar shows the difference in topic proportions between incidental page views and proactive page views.

Notes

1. Based on Alexa top sites rankings, Wikipedia has consistently ranking among the top 10 most trafficked websites globally over the past decade (https://www.alexa.com/topsites).
2. See https://stats.wikimedia.org/EN/TablesWikipediaZZ.htm (accessed December 21, 2018).
3. For example, a 2005 study in *Nature* found Wikipedia content to be almost as reliable as that of Encyclopedia Britannica (21), and a 2014 study in *PLoS one* found that Wikipedia content related to pharmacology was 99.7% accurate compared to a pharmacology textbook, and over 80% complete (22).
4. The Great Firewall of China blocks foreign websites from Chinese IP addresses, including a wide range of websites and social media sites (for a list of blocked websites see GreatFire.org).
5. See https://en.greatfire.org/search/wikipedia-pages (accessed December 31, 2017) for blocked Wikipedia sites.
6. See Welinder, Yana, Victoria Baranetsky, and Brandon Black. "Securing Access to Wikimedia Sites with HTTPS," *Wikimedia Blog*, June 12, 2015, https://blog.wikimedia.org/2015/06/12/securing-wikimedia-sites-with-https/ (accessed December 31, 2017). Oberhaus, Daniel, "Wikipedia's Switch to HTTPS Has Successfully Fought Government Censorship," *Motherboard*, May 26, 2017. https://bit.ly/2T5aEWm (accessed December 31, 2017).
7. Smith, Charlie. "We Had Our Arguments, But We Will Miss You Wikipedia," *Huffington Post*, https://bit.ly/2Ra6AXm (accessed December 31, 2017). "Censorship of Wikipedia," *Wikipedia*. https://bit.ly/2QGQEwp (accessed December 31, 2017).
8. See https://dumps.wikimedia.org/other/pagecounts-raw/.
9. See (2) for plots of total daily page views during this week.
10. Detailed information on the geolocation of users is frequently collected by for-profit internet companies that can monetize this type of data. However, Wikipedia is hosted by the Wikimedia Foundation, a US-based non-profit organization that does not sell user information. Based on our conversations with the Wikimedia Foundation, historical data on page views by geography for the regions we are interested in studying is not available.
11. We also merge together simplified and traditional pages because they redirect to one another. This is the subset of pages that we use for all of our remaining analyses.
12. The homepage is not the only page that provides these curated sets of links. Other such pages include topical and temporal indexes.
13. For links of links of links from the homepage, we only used links of those links of links that were viewed during this time period in order to count only page views that were possibly generated initially from the homepage.
14. See https://wikimedia.org/api/rest_v1/.
15. Another limitation of our method is we cannot adjust for hourly differences between 2-3PM and 4-5PM. We assume that the types of pages viewed between these two hours is not systematically different.
16. An alternative way of approaching this would be to use Wikipedia categories instead of topics. However, Wikipedia categories themselves number in the tens of thousands and are organized in a tree structure. The topic model has the advantage of obtaining clusters specific to the data on hand, which may not be reflected in the Wikipedia categories.
17. We used the Python package Wikipedia-API to do this.
18. For a different metric, we also report the estimated proportion of page views due to browsing by topic in the Appendix

References

[1] Chen Y and Yang DY. The impact of media censorship: 1984 or brave new world? *American Economic Review* 2019; 109(6).

[2] Roberts ME. *Censored: Distraction and Diversion Inside China's Great Firewall*. Princeton: Princeton University Press, 2018.

[3] Enikolopov R, Petrova M and Zhuravskaya E. Media and political persuasion: Evidence from russia. *American Economic Review* 2011; 101(7): 3253–3285.

[4] Pierskalla FM Jan Hand Hollenbach. Technology and collective action: The effect of cell phone coverage on political violence in africa. *American Political Science Review* 2013; 107(2): 207–224. DOI:10.1017/S0003055413000075.

[5] Edmond C. Information manipulation, coordination, and regime change. *The Review of Economic Studies* 2013; 80(4): 1422–1458.

[6] Kalathil S and Boas TC. *Open Networks, Closed Regimes: The Impact of the Internet on Authoritarian Rule*. Washington, D.C.: Carnegie Endowment for International Peace, 2010.

[7] Morozov E. *The Net Delusion: The Dark Side of Internet Freedom*. New York: PublicAffairs, 2011.

[8] MacKinnon R. *Consent of the Networked: The Worldwide Struggle For Internet Freedom*. New York: Basic Books, 2012.

[9] Rød EG and Weidmann NB. Empowering activists or autocrats? the internet in authoritarian regimes. *Journal of Peace Research* 2015; 52(3): 338–351.

[10] Lessig L. *Code: And Other Laws of Cyberspace*. New York: Basic Books, 1999.

[11] Zuckerman E. Cute cats to the rescue? participatory media and political expression. In Allen D and Light JS (eds.) *From Voice to Influence: Understanding Citizenship in a Digital Age*. Chicago: University of Chicago Press, 2015.

[12] Hobbs WR and Roberts ME. How sudden censorship can increase access to information. *American Political Science Review* 2018; 112(3): 621–636.

[13] Shirk SL. *Changing Media, Changing China*. Oxford: Oxford University Press, 2011.

[14] Stockmann D. *Media Commercialization and Authoritarian Rule in China*. Cambridge: Cambridge University Press, 2012.

[15] Blei DM, Ng AY and Jordan MI. Latent dirichlet allocation. *Journal of machine Learning research* 2003; 3(Jan): 993–1022.

[16] Blei DM and Lafferty JD. Dynamic topic models. In *Proceedings of the 23rd international conference on Machine learning*. ACM, pp. 113–120.

[17] Grimmer J. A bayesian hierarchical topic model for political texts: Measuring expressed agendas in senate press releases. *Political Analysis* 2010; 18(1): 1–35.

[18] Quinn KM, Monroe BL, Colaresi M et al. How to analyze political attention with minimal assumptions and costs. *American Journal of Political Science* 2010; 54(1): 209–228.

[19] Roberts ME, Stewart BM, Tingley D et al. Structural topic models for open-ended survey responses. *American Journal of Political Science* 2014; 58(4): 1064–1082.

[20] Lee M and Mimno D. Low-dimensional embeddings for interpretable anchor-based topic inference. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, pp. 1319–1328.

[21] Giles J. Internet encyclopaedias go head to head. *Nature* 2005; (438): 900–901.

[22] Kräenbring J, Penza TM, Gutmann J et al. Accuracy and completeness of drug information in wikipedia: a comparison with standard textbooks of pharmacology. *PloS one* 2014; 9(9): e106930.

Appendix

Page Name	Direct Link to Homepage	Indirect Link to Homepage	Page View Decrease
Wikipedia:首页	0	1	-5598
陶渊明	0	1	-752
Wikipedia:分类索引	0	1	-476
中华人民共和国	0	1	-381
Portal:特色内容	0	0	-379
虚拟化	0	0	-376
英语	0	1	-342
Wikipedia:可供查证	0	1	-332
Wikipedia:关于	0	0	-312
Special:页面分类	0	0	-298
Portal:新闻动态	0	1	-291
Wikipedia:特色条目	0	1	-260
Wikipedia:列明来源	0	1	-257
河北省地区生产总值	0	1	-247
Special:最近更改	0	0	-242
草榴社区	0	0	-238
周迪	1	0	-237
AV女优列表	0	0	-235
碳酸钙	0	1	-232
美国	0	1	-218
Help:目录	0	1	-212
Wikipedia:可靠来源	0	1	-209
Wikipedia:社群首页	0	0	-207
File:Tango-nosources.svg	0	0	-205
Wikipedia:方针与指引	0	1	-203
Wikipedia:互助客栈	0	1	-199
Wikipedia:IRC聊天频道/IRC	0	0	-188
Wikipedia:知识问答	0	0	-188
国际饭店	1	0	-183
Running_Man	0	0	-178
Wikipedia:联络我们	0	0	-177
Wikipedia:字词转换	0	0	-176
5月19日	0	1	-170
台湾	0	1	-169

Page Name	Direct Link to Homepage	Indirect Link to Homepage	Page View Decrease
中国	1	0	-156
细川尚春	1	0	-152
克里斯蒂安·斯特赖希	1	0	-148
Special:Search	0	0	-145
陈用彩	0	0	-144
Wikipedia:上传	0	0	-141
苏珊·安东尼银元	0	1	-139
1美元硬币	1	0	-138
游说集团	1	0	-138
Special:特殊页面	0	0	-137
Special:统计	0	0	-137
苏珊·安东尼	1	0	-137
Wikipedia:欢迎	0	0	-134
美国铸币局	1	0	-133
Wikipedia:特色条目/存档	0	1	-132
德国足球年度最佳教练	1	0	-132
Wikipedia:什么是条目	0	1	-131
Wikipedia:新手入门/主页	0	0	-131
新西兰	0	1	-131
Wikipedia:沙盒	0	0	-130
Wikipedia:特色条目候选	0	1	-130
Wikipedia:VPA	0	0	-129
Portal:首页	0	1	-128
Wikipedia:联系我们/捐款	0	0	-128
自由内容	0	1	-128
自由女神	1	0	-127
Wikipedia:特色列表	0	1	-126
香港	0	1	-124
计算机科学家	0	0	-119
中华民国	0	1	-112
普林斯顿大学诺贝尔奖得主列表	0	1	-111
香港博物馆列表	0	1	-110
铃木一朗	0	1	-109
武媚娘传奇	0	1	-106

	Direct Link	Indirect Link	Page View
Page Name	to Homepage	to Homepage	Decrease
圣路易斯华盛顿大学诺贝尔奖得主列表	0	1	-105
俄罗斯	0	1	-101
英国	0	1	-97
成县	0	1	-95
Template:Fact	0	1	-92
奇怪的保姆	0	0	-88
武则天	0	1	-85
Wikipedia:免责声明	0	0	-84
Wikipedia:小作品	0	1	-84
第二次世界大战	0	1	-80
新加坡	0	1	-79
爱·回家_(电视剧)	0	0	-77
李承乾	0	1	-74
法国	0	1	-74
日语	0	1	-73
未知艺术家	0	0	-72
未知艺术家_(artist)	0	0	-72
通用串行总线	0	0	-71
Special:最新页面	0	0	-70
鸡奸	0	1	-70
唐高宗	0	1	-68
Who_Are_You－学校2015	0	0	-67
习近平	0	1	-67
日语假名	0	1	-67
慕尼黑	0	1	-66
中国大陆	0	1	-64
3000安打俱乐部	0	0	-63
中国人民解放军海军	0	1	-63
姚文智	0	0	-63
平文式罗马字	0	1	-63
AV女优	0	1	-62
金正恩	0	1	-62

Figure 4. Pages most affected by censorship, May 19th 4-5PM - May 19th 2-3PM. Direct link to homepage is whether or not the page was linked directly from the Wikipedia homepage on May 19th. Indirect link to homepage is whether or not the page was linked to a link or linked to a link of a link from the Wikipedia homepage on May 19th.

Table 2. Topics Estimated By Topic Model

Label	Translated Highest Probability Words	Est. Mainland Views/Hr	Est. Prop Browsing
Video Games	Games, games, series, launch, player, platform, translation	1744	0.08
Animation and Comics	Works, animation, publishing, comics, Japanese, girls, serial	2943	0.09
Japanese celebrities	Born, female, born, real name, male, Japanese, av	2866	0.10
TV Shows	Show, TV, host, host, radio, wireless TV, production	2255	0.10
TV Series	Starring, TV series, airing, story, premiere, English, English	2655	0.10
Movies	Movie, release, english, starring, director, film, film	2826	0.11
Animated Movies	Hero, animation, series, english, disney, sci-fi, comics	1321	0.11
Korean Pop	Korea, Korean, Korean, members, debut, combination, group	3392	0.12
Chinese TV shows	Broadcast, play, TV, show, broadcast, TV station, first	1798	0.12
Wikipedia Entries	Utc, entry, article, wiki, wikipedia, hope, message	2839	0.13
Chinese Novels	Series, character, story, character, protagonist, novel, appearance	2336	0.13
Software programs	Software, programs, systems, design, files, languages, users	2072	0.13
Computer Systems	System, program, technology, computer, windows, can, function	2792	0.14
Music	Music, singer, song, album, record, release, popular	3103	0.15
Actors and Actresses	Actor, performance, movie, best, protagonist, play, nomination	2989	0.16
Information Technology	System, information, technology, data, network, standard, computer	2278	0.19
Hong Kong Actors and Actresses	Hong Kong, English, English, special, age, miss, executive	2943	0.19
Math	Displaystyle, function, representation, can, equation, space, math	2250	0.20
Weapons	Weapon, tank, millimeter, pacific, design, production, launch	1486	0.20
Japan	Japan, Japanese, Tokyo, Japanese, Meiji, one, Hokkaido	3194	0.21
Music, Movie, and TV Awards	Film, actor, get, music, album, release, director	2612	0.21
Business and Stock Market	Company, limited, group, shares, established, affiliated, listed	2668	0.22
Diseases	Disease, possible, patient, infection, virus, cause, symptoms	2700	0.23
Animals	Animals, creatures, discoveries, oceans, scientific names, them, species	2258	0.24
Consumer Products	Brand, product, market, industry, technology, automotive, global	1778	0.24
U.S. and U.K.	America, English, UK, first, age, one, become	2614	0.24
Soccer	Football,champion,spain,professional,team,effective,brazilian	3008	0.24
Biology	Biology, cells, genes, proteins, effects, food, food	1937	0.24
Locations in Hong Kong and Macau	Located, Macau, architecture, Hong Kong, center, park, hotel	2151	0.25
Transportation	Station, highway, road, system, located, railway, km	1639	0.26
Logic	Different, can, usually, generally, called, for example, therefore	7260	0.27
Aviation	Aviation,airport,airport,airplane,airborne,built,code	2142	0.27
Categories	Including, other, part, of which, all, and, in addition to	3183	0.27
Taiwan	Taiwan, Taiwan, Republic of China, Taipei, North City, located in Kaohsiung City	2966	0.27
Physics	Physics, particle, theory, energy, action, can, direction	2046	0.27
Public Transit	Service, public, offer, bus, traffic, tunnel, route	915	0.28
Written Languages	Use, English, representation, letters, text, Chinese, representative	2624	0.28
General Words	Mainly, one, taken from, currently, initially, to date, named	387	0.28
General Words II	Yes, different, called, this, due, some, because	5661	0.28
Chemistry	Reaction, material, chemistry, element, molecule, compound, atom	2735	0.28
U.S., U.K., Australia, Canada	United States, English, United Kingdom, Canada, Royal, Australia, London	2604	0.29

Tragic Events and Crimes	Event,occurrence,court,death,cause,police,crime	2303	0.29
Plants	Plant, scientific name, animal, distribution, species, region, forest	2388	0.29
Military	Navy, liberation army, troops, missiles, military, army, combat	2101	0.29
Space and Astronomy	Earth, planet, space, universe, rocket, galaxy, star	1520	0.30
Country Statistics and Domain Names	World, country, Europe, international, Singapore, region, global	2903	0.30
Banks	Group, corporate, international, service, market, product, bank	2204	0.30
Taiwan Politics	Republic of China, politics, figures, middle school, current, member, born in	2656	0.31
Culture and Society	Culture, history, development, influence, modern, society, tradition	3310	0.31
Dialects	India, nationality, language, language, pinyin, official, dialect	2554	0.32
Research	Research, science, theory, method, field, analysis, presentation	2150	0.32
Labor and the Economy	Bank, management, work, government, institution, economy, business	2495	0.33
English Novels	Famous, works, writer, novel, one, character, art	2870	0.33
Education and University	Education, university, school, college, ranking, research, Sweden	1788	0.33
History of the Military	–, war, second, first, military, battle, troops	2775	0.33
International Sports	Football, nba, championship, league, club, match, athlete	2200	0.34
East Asia	China, Vietnam, North Korea, China, one, region, ancient	2041	0.34
Organization of Societies	Society,ism,sports,activity,believes,organization,law	2461	0.34
Materials	Use, can, cause, speed, increase, glass, form	2051	0.34
Regions Statistics	City, located, population, region, capital, center, place	3998	0.35
Ancient History	Century, history, period, egypt, times, BC, ancient	2617	0.36
Religion	Religion, Buddhism, Christianity, Faith, Catholicism, One, Christ	2373	0.36
European History	Italy, dynasty, empire, king, monarch, rome, german	2352	0.38
China Scenic Areas	China, Beijing, China, Shanghai, Zhejiang, Scenic Area, Taiwan	3199	0.38
Democracy	Government, politics, freedom, democracy, policy, institution, revolution	3623	0.38
Highways and High Speed Railway	Railway, high speed, kilometers, traffic, highway, train, subway	1862	0.39
Japanese History	Age, period, emperor, after, later, father, son	3293	0.39
French Geography	Population, french, french, area, square kilometers, paris, de	2488	0.39
Economic Development	Economy, development, country, production, region, trade, world	2567	0.39
Soviet History	State, federation, russia, soviet, president, organization, government	2986	0.40
Government and Legal System of China	People, China, Republic, Government, Administration, Republic of China, Republic	3106	0.40
World Military History	War, Germany, period, end, start, happen, become	3065	0.41
Education, Professors and Students	University, college, school, student, engineering, education, professor	1816	0.41
Ming and Qing Dynasty History	Qing Dynasty, Ming Dynasty, Mongolia, Emperor, thirteen, minister, Qing Dynasty	2875	0.42
Chinese Communist Party	Central, CCP, committee, Communist Party, committee, representative, chairman	2718	0.42

Geography	Located, area, kilometers, area, north, south, square kilometers	3209	0.43
Tang Dynasty History	Emperor, incumbent, Tang Dynasty, Queen, Princess, period, monarch	4881	0.48
International Heritage and Cultural Sites	City, center, architecture, located, culture, history, one	3017	0.48