SMaPP Global Special Issue - Original Research

SAGE Open
January-March 2020: 1–14
© The Author(s) 2020
DOI: 10.1177/2158244019894068
journals.sagepub.com/home/sgo

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 such as 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 more than 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 article, 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 that 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. Furthermore, 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 (Chen & Yang, 2019; Roberts, 2018) that augments studies of the effects of selective censorship of political information (Edmond, 2013; Enikolopov et al., 2011; Kalathil & Boas, 2010; Lessig, 1999; MacKinnon, 2012; Morozov, 2011; Pierskalla & Hollenbach, 2013; Rød & Weidmann, 2015) and the unintended consequences of entertainment-related censorship (Hobbs & Roberts, 2018; Zuckerman, 2015).

We construct a timeline of page views for each of 372,208 pages on Chinese language Wikipedia that allows us to

[1]Stanford University, CA, USA
[2]University of California, San Diego, La Jolla, USA

Corresponding Author:
Margaret E. Roberts, Associate Professor, Department of Political Science, University of California, San Diego, 9500 Gilman Dr, #0521, La Jolla, CA 92093, USA.
Email: meroberts@ucsd.edu

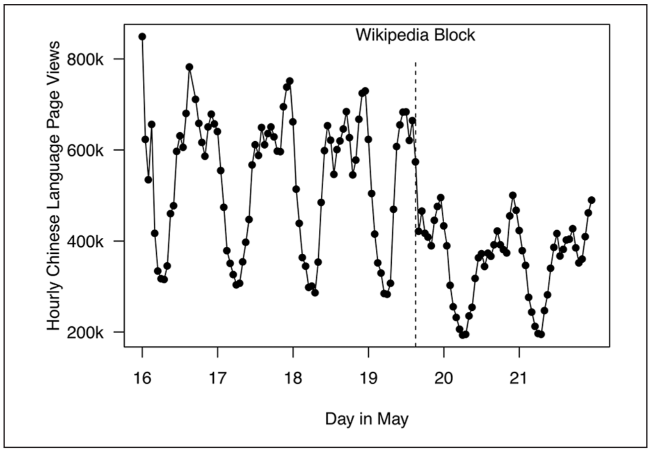


Figure 1. Page views of Chinese language Wikipedia by hour, May 16 to 21, 2015.
Note. The Wikipedia block occurred during the afternoon of May 19, 2015.

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. Although our data do not allow us to differentiate directly between mainland and nonmainland 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 section "Homepage Effects on Page Views," 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 nonmainland viewers. The section "Topical Differences in Proactive Versus Incidental Consumption" uses a topic model to describe the types of pages that were viewed by mainland users incidentally versus proactively, and the last section concludes the article.

Overall Impact of Wikipedia Censorship in China

Chinese language Wikipedia was launched in 2001. Although media in China is tightly controlled (Shirk, 2011; Stockmann, 2012), 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 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]

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 unencrypted Hyper Text Transfer Protocol (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 show that the block occurred sometime within the hour of 3:00 p.m. 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 1 a.m. and 5 a.m. China time) and then rises during the day and evening. This pattern is interrupted midday—between 3 p.m. and 4 p.m. 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] Although 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 (VPN), this number is likely small for two reasons. First, there are few VPN users in mainland China (Chen & Yang, 2019; Roberts, 2018), 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 (Roberts, 2018).

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 2 p.m. and 3 p.m. on May 19, directly before the block. In comparison, there were 421,663 total page views between 4 p.m. and 5 p.m. on May 19, resulting in a 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 in 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 2 days in the 2 p.m. to 3 p.m. hour the week prior to the block. We do this to exclude many pages in the Wikipedia data set that are blank pages, which are likely to be viewed less than twice 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 (2–3 p.m. on May 19) and after the block (4–5 p.m. on May 19) and selected the 100 pages with the largest page view decreases. Although the top pages only provide a partial picture of the Wikipedia content affected by the block, the 100 pages (0.026% of all pages) that lost the most views account for more than 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 A1).

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 3 p.m. and 5 p.m., and was the 47th most affected page in the data set. "Ghost marriage" a traditional custom lost 196 views, the 25th most affected page in the data set. 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.

To systematically estimate the impact of the homepage on mainland views on Wikipedia, we document all pages in our data set that were linked to the homepage on May 19. 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:历史上的今天/5月, as well as an archive of facts of newly created Wikipedia pages from the "Did you know section": Wikipedia:新条目推荐/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

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.

Type of page	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

Note. Pages that experienced the largest average decrease in page views on average were those with links to the homepage.

use the Wikipedia application programming interface (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 to 3 p.m. and 4 to 5 p.m. 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 51% 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.

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 provides 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 were 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. Although these data are 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, although 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 cannot 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 its appearance on the home page—which we will call *proactive* page views.

We use the example of the Chinese language page for the US$1 coin (1美元硬币) 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 US$1 coin was not a very popular page on May 18, with total page views hovering around 10 per hour the day before it was featured on the homepage. However on May 19, the page was featured on the homepage and was viewed a whole lot more—150 times in the 2 to 3 p.m. 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 19 in the 2 to 3 p.m. hour $(V_{i,2\text{p.m.},5/19})$ and May 18 in the 2 to 3 p.m. hour $(V_{i,2\text{p.m.},5/18})$ is an estimate of the total increase in page views for the US$1 coin page because of the homepage—page views driven by incidental consumption—for the 2 to 3 p.m. hour $(V_{\text{incidental}})$, as shown in Equation 1. (Note: Because views cannot be negative, we take the maximum of this difference and zero.)

$$V_{i,\text{incidental}} = \max\left(V_{i,2\text{p.m.},5/19} - V_{i,2\text{p.m.},5/18}, 0\right) \tag{1}$$

Because we cannot distinguish before the block between mainland and nonmainland viewers, this difference, $V_{i,\text{incidental}}$, includes incidental consumption from both mainland China users (m) and nonmainland China users (m'). Sometime during the 3 to 4 p.m. hour, Chinese Wikipedia page views from mainland China were blocked, leading to a large decrease in page views of the US$1 coin page between the hours of 2 to 3 p.m. and 4 to 5 p.m., when page views dipped to around 25 views. Page views further decrease on May 20 to around five views per hour, when Chinese language Wikipedia is still blocked, and the US$1 coin is no longer featured on the Wikipedia homepage. Because the block only affected mainland Chinese users, the difference between the 4 to 5 p.m. page views on May 19 $(V_{i,4\text{p.m.},5/19})$ and 4 to 5 p.m. page views on May 20 $(V_{i,4\text{ p.m.},5/20})$ is an estimate of the incidental consumption generated by the homepage for nonmainland viewers only $(V_{i,\text{incidental},m'})$. Nonmainland incidental consumption of pages linked directly or indirectly to the homepage can then be estimated as

$$V_{i,\text{incidental},m'} = \max(V_{i,4\text{p.m.},5/19} - V_{i,4\text{p.m.},5/20}, 0) \tag{2}$$

Using the estimates from Equations 1 and 2, we can then obtain an estimate for incidental consumption from mainland China. Incidental consumption from mainland China is the difference between $V_{i,\text{incidental}}$ and $V_{i,\text{incidental},m'}$ for each page linked directly or indirectly to the homepage.

$$V_{i,\text{incidental},m} = \max(V_{i,\text{incidental}} - V_{i,\text{incidental},m'}, 0) \tag{3}$$

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

$$V_{i,\text{total},m} = \max(V_{i,2\text{ p.m.},5/19} - V_{i,4\text{p.m.},5/19}, 0) \tag{4}$$

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

$$V_{i,\text{incidental},m} = \min(V_{i,\text{total},m}, V_{i,\text{incidental},m}) \tag{5}$$

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

$$V_{i,\text{proactive},m} = V_{i,\text{total},m} - V_{i,\text{incidental},m} \tag{6}$$

Altogether, we can roughly estimate the total page views of each page coming 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. Because 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 approximately 42% of the loss in page views due to the block were page views based on incidental consumption—views above and beyond what the page generally received on the day it was linked to the homepage. This indicates that although censorship did have an impact on limiting the information users were proactively looking for, its impact on incidental consumption of information was also very important, accounting for almost half of total page views.

Topical Differences in Proactive Versus Incidental Consumption

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

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 (Blei et al., 2003; Blei and Lafferty, 2006; Grimmer, 2010; Quinn et al., 2010). 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 data set. The summary information for each Wikipedia page is the first paragraph of the page and exists for the majority of content 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 data set that were viewed in the hour directly before or after the May 19 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 (Roberts et al., 2014). We used an automated method to select the number of topics (Lee & Mimno, 2014). 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 A1), 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 ($V_{i,\text{total},m}$). Outside of general word topics, the most popular topics in mainland China include many topics about history and politics, including the Chinese Communist Party, 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 $i = 1,\ldots,N,$ 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 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 $V_{i,\text{incidental},m}$. 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 $V_{i,\text{proactive},m}$. 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:

$$\vec{T}_{\text{incidental}} = \frac{\sum_{i=1}^{N} \vec{\theta}_i \times V_{i,\text{incidental},m}}{\sum_{i=1}^{N} V_{i,\text{incidental},m}}$$

and the proportion of proactive page views in each topic:

$$\vec{T}_{\text{proactive}} = \frac{\sum_{i=1}^{N} \vec{\theta}_i \times V_{i,\text{proactive},m}}{\sum_{i=1}^{N} V_{i,\text{proactive},m}}$$

We take the difference $\vec{T}_{\text{incidental}} - \vec{T}_{\text{proactive}}$ 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 Japanese celebrities, animated movies, music, and sports. Incidental page views, however, were disproportionately focused on pages that talked about politics and history, ranging from 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 were 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 such as 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 straightforward for users to seek out information that they queried proactively on Wikipedia on other websites. However, users would not know what information they would have arrived at on Chinese language Wikipedia incidentally, and as a result this information would be very difficult to seek out from other websites without Wikipedia. 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.

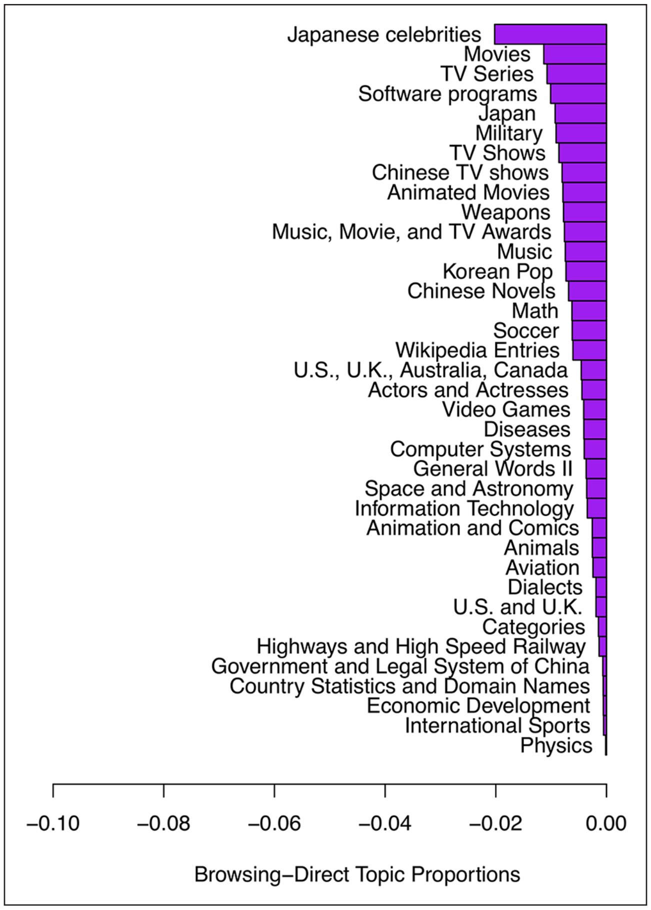


Figure 2. Topics associated with proactive page views, estimated from the structural topic model.
Note. The length of the bar shows the difference in topic proportions between incidental page views and proactive page views.

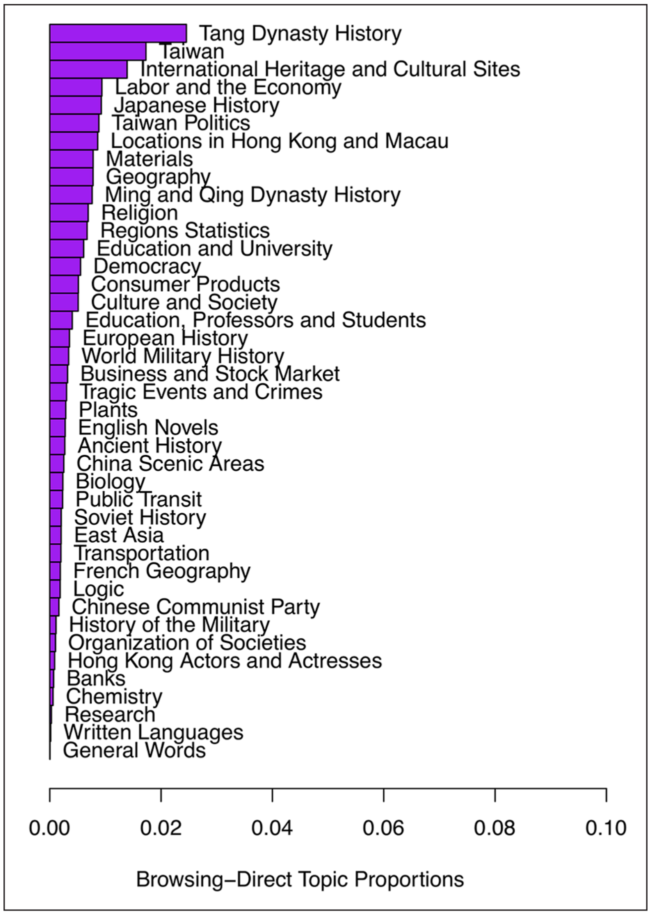


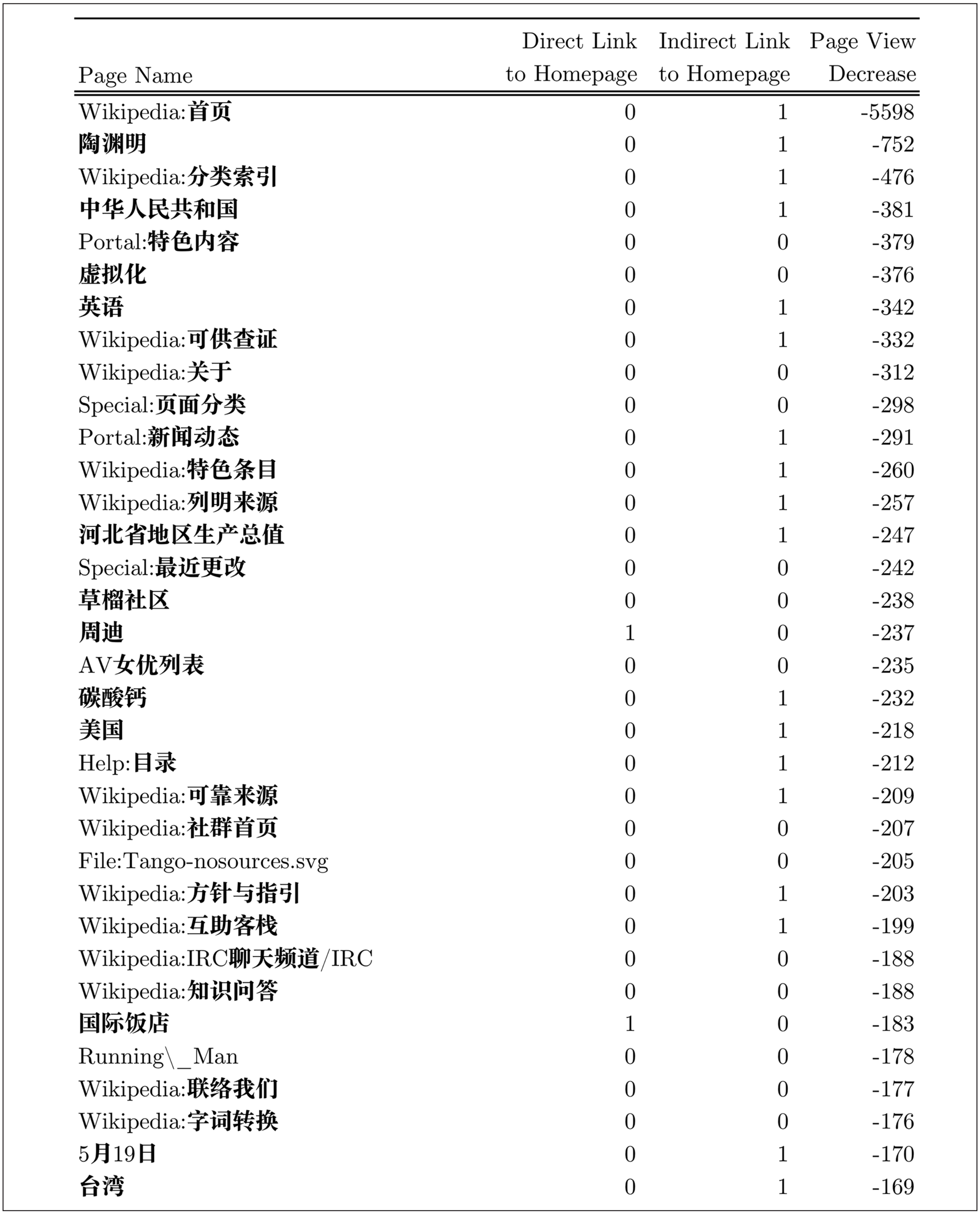
Figure 3. Topics associated with incidental page views, estimated from the structural topic model.
Note. The length of the bar shows the difference in topic proportions between incidental page views and proactive page views.

Conclusion

In this article, 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 professionals needing 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 overcome the costs of censorship to seek out information (Chen & Yang, 2019; Roberts, 2018), greater impediments to exploration may endogenously decrease a proactive demand for certain types of information because users are less likely to come across a greater variety of information incidentally.

Appendix

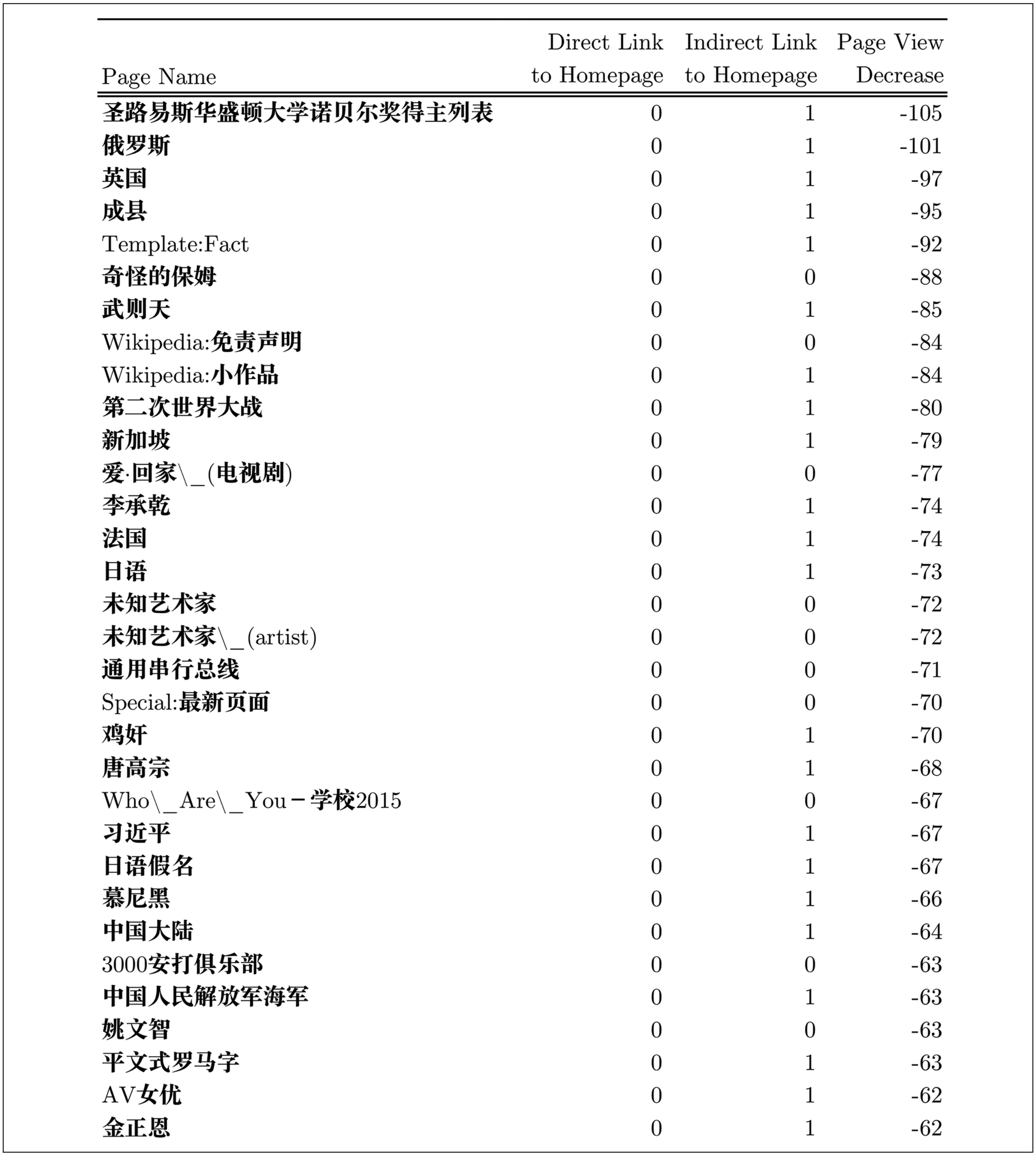


Page Name	Direct Link to Homepage	Indirect Link to Homepage	Page View Decrease
Wikipedia:首页	0	1	-5598
陶渊明	0	1	-752
Wikipedia:分类索引	0	1	-476
中华人民共和国	0	1	-381
Portal:特色内容	0	0	-379
虚拟化	0	0	-376
英语	0	1	-342
Wikipedia:可供查证	0	1	-332
Wikipedia:关于	0	0	-312
Special:页面分类	0	0	-298
Portal:新闻动态	0	1	-291
Wikipedia: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

(continued)

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

(continued)



Page Name	Direct Link to Homepage	Indirect Link to Homepage	Page View 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 A1. Pages most affected by censorship, May 19, 4 to 5 p.m. − May 19, 2 to 3 p.m.
Note. Direct link to homepage is whether or not the page was linked directly from the Wikipedia homepage on May 19. 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 19.

Table A1. Topics Estimated by Topic Model.

Label	Translated highest probability words	Estimated mainland views/hour	Estimated proportion browsing
Video games	Games, games, series, launch, player, platform, translation	1,744	.08
Animation and comics	Works, animation, publishing, comics, Japanese, girls, serial	2,943	.09
Japanese celebrities	Born, female, born, real name, male, Japanese, av	2,866	.10
TV shows	Show, TV, host, host, radio, wireless TV, production	2,255	.10
TV series	Starring, TV series, airing, story, premiere, English	2,655	.10
Movies	Movie, release, English, starring, director, film	2,826	.11
Animated movies	Hero, animation, series, English, Disney, sci-fi, comics	1,321	.11
Korean pop	Korea, Korean, members, debut, combination, group	3,392	.12
Chinese TV shows	Broadcast, play, TV, show, TV station, first	1,798	.12
Wikipedia entries	Utc, entry, article, wiki, Wikipedia, hope, message	2,839	.13
Chinese novels	Series, character, story, protagonist, novel, appearance	2,336	.13
Software programs	Software, programs, systems, design, files, languages, users	2,072	.13
Computer systems	System, program, technology, computer, windows, can, function	2,792	.14
Music	Music, singer, song, album, record, release, popular	3,103	.15
Actors and actresses	Actor, performance, movie, best, protagonist, play, nomination	2,989	.16
Information technology	System, information, technology, data, network, standard, computer	2,278	.19
Hong Kong actors and actresses	Hong Kong, English, special, age, miss, executive	2,943	.19
Math	Display style, function, representation, can, equation, space, math	2,250	.20
Weapons	Weapon, tank, millimeter, pacific, design, production, launch	1,486	.20
Japan	Japan, Japanese, Tokyo, Meiji, one, Hokkaido	3,194	.21
Music, movie, and TV awards	Film, actor, get, music, album, release, director	2,612	.21
Business and stock market	Company, limited, group, shares, established, affiliated, listed	2,668	.22
Diseases	Disease, possible, patient, infection, virus, cause, symptoms	2,700	.23
Animals	Animals, creatures, discoveries, oceans, scientific names, them, species	2,258	.24
Consumer products	Brand, product, market, industry, technology, automotive, global	1,778	.24
U.S. and U.K.	America, English, UK, first, age, one, become	2,614	.24
Soccer	Football, champion, Spain, professional, team, effective, Brazilian	3,008	.24
Biology	Biology, cells, genes, proteins, effects, food	1,937	.24
Locations in Hong Kong and Macau	Located, Macau, architecture, Hong Kong, center, park, hotel	2,151	.25
Transportation	Station, highway, road, system, located, railway, km	1,639	.26
Logic	Different, can, usually, generally, called, for example, therefore	7,260	.27
Aviation	Aviation, airport, airplane, airborne, built, code	2,142	.27
Categories	Including, other, part, of which, all, and, in addition to	3,183	.27
Taiwan	Taiwan, Republic of China, Taipei, North City, located in Kaohsiung City	2,966	.27

(continued)

Table A1. (continued)

Label	Translated highest probability words	Estimated mainland views/hour	Estimated proportion browsing
Physics	Physics, particle, theory, energy, action, can, direction	2,046	.27
Public transit	Service, public, offer, bus, traffic, tunnel, route	915	.28
Written languages	Use, English, representation, letters, text, Chinese, representative	2,624	.28
General words	Mainly, one, taken from, currently, initially, to date, named	387	.28
General words II	Yes, different, called, this, due, some, because	5,661	.28
Chemistry	Reaction, material, chemistry, element, molecule, compound, atom	2,735	.28
U.S., U.K., Australia, Canada	United States, English, United Kingdom, Canada, Royal, Australia, London	2,604	.29
Tragic events and crimes	Event, occurrence, court, death, cause, police, crime	2,303	.29
Plants	Plant, scientific name, animal, distribution, species, region, forest	2,388	.29
Military	Navy, liberation army, troops, missiles, military, army, combat	2,101	.29
Space and astronomy	Earth, planet, space, universe, rocket, galaxy, star	1,520	.30
Country statistics and domain names	World, country, Europe, international, Singapore, region, global	2,903	.30
Banks	Group, corporate, international, service, market, product, bank	2,204	.30
Taiwan politics	Republic of China, politics, figures, middle school, current, member, born in	2,656	.31
Culture and society	Culture, history, development, influence, modern, society, tradition	3,310	.31
Dialects	India, nationality, language, pinyin, official, dialect	2,554	.32
Research	Research, science, theory, method, field, analysis, presentation	2,150	.32
Labor and the economy	Bank, management, work, government, institution, economy, business	2,495	.33
English novels	Famous, works, writer, novel, one, character, art	2,870	.33
Education and university	Education, university, school, college, ranking, research, Sweden	1,788	.33
History of the military	—, war, second, first, military, battle, troops	2,775	.33
International sports	Football, NBA, championship, league, club, match, athlete	2,200	.34
East Asia	China, Vietnam, North Korea, one, region, ancient	2,041	.34
Organization of societies	Society, ism, sports, activity, believes, organization, law	2,461	.34
Materials	Use, can, cause, speed, increase, glass, form	2,051	.34
Regions statistics	City, located, population, region, capital, center, place	3,998	.35
Ancient history	Century, history, period, Egypt, times, BC, ancient	2,617	.36
Religion	Religion, Buddhism, Christianity, Faith, Catholicism, One, Christ	2,373	.36
European history	Italy, dynasty, empire, king, monarch, Rome, German	2,352	.38
China scenic areas	China, Beijing, Shanghai, Zhejiang, Scenic Area, Taiwan	3,199	.38
Democracy	Government, politics, freedom, democracy, policy, institution, revolution	3,623	.38
Highways and high speed railway	Railway, high speed, kilometers, traffic, highway, train, subway	1,862	.39
Japanese history	Age, period, emperor, after, later, father, son	3,293	.39

(continued)

Table A1. (continued)

Label	Translated highest probability words	Estimated mainland views/hour	Estimated proportion browsing
French geography	Population, French, French, area, square kilometers, Paris, de	2,488	.39
Economic development	Economy, development, country, production, region, trade, world	2,567	.39
Soviet history	State, federation, Russia, Soviet, president, organization, government	2,986	.40
Government and legal system of China	People, China, Republic, government, administration, Republic of China	3,106	.40
World military history	War, Germany, period, end, start, happen, become	3,065	.41
Education, professors, and students	University, college, school, student, engineering, education, professor	1,816	.41
Ming and Qing dynasty history	Qing Dynasty, Ming Dynasty, Mongolia, Emperor, thirteen, minister, Qing Dynasty	2,875	.42
Chinese Communist Party	Central, CCP, committee, Communist Party, representative, chairman	2,718	.42
Geography	Located, area, kilometers, area, north, south, square kilometers	3,209	.43
Tang Dynasty history	Emperor, incumbent, Tang Dynasty, Queen, Princess, period, monarch	4,881	.48
International heritage and cultural sites	City, center, architecture, located, culture, history, one	3,017	.48

Note. NBA = National Basketball Association.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iD

Margaret E. Roberts https://orcid.org/0000-0001-6900-4366

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
3. For example, a 2005 study in *Nature* found Wikipedia content to be almost as reliable as that of Encyclopedia Britannica (Giles, 2005), and a 2014 study in *PLOS ONE* found that Wikipedia content related to pharmacology was 99.7% accurate compared with a pharmacology textbook, and more than 80% complete (Kräenbring et al., 2014).
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 Great Fire.org).
5. See https://en.greatfire.org/search/wikipedia-pages, for blocked Wikipedia sites.
6. See Welinder et al. (2015) and Oberhaus (2017).
7. Smith (2017); "Censorship of Wikipedia," *Wikipedia*. https://bit.ly/2QGQEwp.
8. See https://dumps.wikimedia.org/other/pagecounts-raw/
9. See Roberts (2018) 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 these types of data. However, Wikipedia is hosted by the Wikimedia Foundation, a U.S.-based nonprofit 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 are 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 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 that we cannot adjust for hourly differences between 2 to 3 p.m. and 4 to 5 p.m. We assume that the types of pages viewed between these 2 hr are 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 at 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

Blei, D. M., & Lafferty, J. D. (2006). Dynamic topic models. In *Proceedings of the 23rd international conference on machine learning* (pp. 113–120). Pittsburgh, PA: Association for Computing Machinery.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, *3*, 993–1022.

Chen, Y., & Yang, D. Y. (2019). The impact of media censorship: 1984 or brave new world? *American Economic Review*, *109*(6), 2294–2332.

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

Enikolopov, R., Petrova, M., & Zhuravskaya, E. (2011). Media and political persuasion: Evidence from Russia. *American Economic Review*, *101*(7), 3253–3285.

Giles, J. (2005). Internet encyclopaedias go head to head. *Nature*, *438*, 900–901.

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

Hobbs, W. R., & Roberts, M. E. (2018). How sudden censorship can increase access to information. *American Political Science Review*, *112*(3), 621–636.

Kalathil, S., & Boas, T. C. (2010). *Open networks, closed regimes: The impact of the Internet on authoritarian rule*. Carnegie Endowment for International Peace.

Kräenbring, J., Penza, T. M., Gutmann, J., Muehlich, S., Zolk, O., Wojnowski, L., . . . Sarikas, A. (2014). Accuracy and completeness of drug information in Wikipedia: A comparison with standard textbooks of pharmacology. *PLOS ONE*, *9*(9), Article e106930.

Lee, M., & Mimno, D. (2014). Low-dimensional embeddings for interpretable anchor-based topic inference. In *Proceedings of the 2014 conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 1319–1328). Doha, Qatar: Association for Computational Linguistics.

Lessig, L. (1999). *Code: And other laws of cyberspace*. Basic Books.

MacKinnon, R. (2012). *Consent of the networked: The worldwide struggle for Internet freedom*. Basic Books.

Morozov, E. (2011). *The net delusion: The dark side of Internet freedom*. PublicAffairs.

Oberhaus, D. (2017, May 26). Wikipedia's switch to HTTPS has successfully fought government censorship. *Motherboard*. https://bit.ly/2T5aEWm

Pierskalla, J. H., & Hollenbach, F. M. (2013). Technology and collective action: The effect of cell phone coverage on political violence in Africa. *American Political Science Review*, *107*(2), 207–224. https://doi.org/10.1017/S0003055413000075

Quinn, K. M., Monroe, B. L., Colaresi, M., Crespin, M. H., & Radev, D. R. (2010). How to analyze political attention with minimal assumptions and costs. *American Journal of Political Science*, *54*(1), 209–228.

Rød, E. G., & Weidmann, N. B. (2015). Empowering activists or autocrats? The Internet in authoritarian regimes. *Journal of Peace Research*, *52*(3), 338–351.

Roberts, M. E. (2018). *Censored: Distraction and diversion inside China's great firewall*. Princeton University Press.

Roberts, M. E., Stewart, B. M., Tingley, D., Lucas, C., Leder-Luis, L., Gadarian, S. K., . . . Rand, D. G. (2014). Structural topic models for open-ended survey responses. *American Journal of Political Science*, *58*(4), 1064–1082.

Shirk, S. L. (2011). *Changing media, changing China*. Oxford University Press.

Smith, C. (2017). We had our arguments, but we will miss you Wikipedia. *Huffington Post*. https://bit.ly/2Ra6AXm

Stockmann, D. (2012). *Media commercialization and authoritarian rule in China*. Cambridge University Press.

Welinder, Y., Victoria, B., & Brandon, B. (2015, June 12). Securing access to Wikimedia sites with HTTPS. *Wikimedia Blog*. https://blog.wikimedia.org/2015/06/12/securing-wikimedia-sites-with-https/

Zuckerman, E. (2015). Cute cats to the rescue? Participatory media and political expression. In D. Allen & J. S. Light (Eds.), *From voice to influence: Understanding citizenship in a digital age* (pp. 131–154). The University of Chicago Press.