Analyzing Accessibility of Wikipedia Projects Around the World

Analyzing Accessibility of Wikipedia Projects Around the World

INTERNET MONITOR

May 2017

Analyzing Accessibility of Wikipedia Projects Around the World

Justin Clark
Robert Faris
Rebekah Heacock Jones

INTERNET MONITOR is a research project to evaluate, describe, and summarize the means, mechanisms, and extent of Internet content controls and Internet activity around the world.

thenetmonitor.org

INTERNET MONITOR is a project of the Berkman Center for Internet & Society. http://cyber.harvard.edu

23 Everett Street • Second Floor • Cambridge, Massachusetts 02138

+1 617.495.7547 • +1 617.495.7641 (fax) • http://cyber.harvard.edu • hello@cyber.harvard.edu

ABSTRACT

This study, conducted by the Internet Monitor project at the Berkman Klein Center for Internet & Society, analyzes the scope of government-sponsored censorship of Wikimedia sites around the world. The study finds that, as of June 2016, China was likely censoring the Chinese language Wikipedia project, and Thailand and Uzbekistan were likely interfering intermittently with specific language projects of Wikipedia as well.

However, considering the widespread use of filtering technologies and the vast coverage of Wikipedia, our study finds that, as of June 2016, there was relatively little censorship of Wikipedia globally. In fact, our study finds there was less censorship in June 2016 than before Wikipedia's transition to HTTPS-only content delivery in June 2015. HTTPS prevents censors from seeing which page a user is viewing, which means censors must choose between blocking the entire site and allowing access to all articles. This finding suggests that the shift to HTTPS has been a good one in terms of ensuring accessibility to knowledge.

The study identifies and documents the blocking of Wikipedia content using two complementary data collection and analysis strategies: a client-side system that collects data from the perspective of users around the globe and a server-side tool to analyze traffic coming in to Wikipedia servers. Both client- and server-side methods detected events that we consider likely related to censorship, in addition to a large number of suspicious events that remain unexplained. The report features results of our data analysis and insights into the state of access to Wikipedia content in 15 select countries.

AUTHORS

Justin Clark is a software developer at the Berkman Klein Center for Internet & Society at Harvard University. Most recently, he has been adapting, designing and crafting systems for mapping the contours of information control on the Internet.

Robert Faris is the Research Director at the Berkman Klein Center for Internet & Society at Harvard University. His recent research has been focused on developing and applying methods for studying the networked public sphere.

Rebekah Heacock Jones is a former senior project manager for the Berkman Klein Center for Internet & Society.

ACKNOWLEDGEMENTS

The authors would like to thank the following people for their helpful contributions and feedback: Grant Baker, Patrick Drown, Urs Gasser, Casey Tilton, Zhou Zhou, and Jonathan Zittrain.

COVER IMAGE
""Greater Chicago Metropolitan Area"
Image Credit: (NASA, International Space Station, 02/02/12)
https://spaceflight.nasa.gov/gallery/images/station/crew-30/html/iss030e062540.html

BERKMAN KLEIN CENTER
FOR INTERNET & SOCIETY AT HARVARD UNIVERSITY

Analyzing Accessibility of Wikipedia Projects Around the World[1]

[1] This report was supported by the Wikimedia Foundation, and the Berkman Klein Center was selected as the research partner in a project to map the accessibility of Wikipedia around the world. The project is listed in Wikimedia's Research wiki here: https://meta.wikimedia.org/wiki/Research:Analyzing_Accessibility_of_Wikipedia_Projects_Around_the_World.

Table of Contents

Introduction

As one of the largest online repositories of user-generated content in the world, covering topics that range from the general reference[3] to the highly controversial,[4] Wikipedia has repeatedly found itself the target of government censors in countries ranging from China to Iran to Uzbekistan. In some cases, individual articles have been singled out: Turkey has blocked a handful of articles related to reproductive biology, as well as at least one political article;[5] in 2008, a number of ISPs in the United Kingdom blocked access to an article about the German band Scorpion's album, "Virgin Killer," the album art for which was a provocative image of a naked child.[6] In other cases, one or two offending articles have prompted wholesale blocks of the site: Russia has intermittently blocked access to all of Wikipedia out of concerns around articles related to the smoking of marijuana;[7] and in 2006, Pakistan temporarily blocked the site in response to an article on "Draw Mohammed Day," which violated certain religious prohibitions against visual depictions of Mohammed.[8] Syria,[9] China,[10] Iran,[11] Tunisia,[12] and Uzbekistan[13] have all blacklisted the site at various times without publicly citing specific content concerns.

A detailed look at the filtering of specific Wikipedia articles can serve as a window into the kinds of content—political, historical, religious, sexual, cultural, drug- or alcohol-related—that trigger censorship in different countries. Censorship of Wikipedia became slightly more complex, however,

---

[3] "Portal:Contents/Categories: General Reference," *Wikipedia*, https://en.wikipedia.org/wiki/Portal:Contents/Categories#General_reference.
[4] "Wikipedia:List of controversial issues," *Wikipedia*, https://en.wikipedia.org/wiki/Wikipedia:List_of_controversial_issues.
[5] "Wikipedia releases warning on Turkey's censorship, monitoring," *Hurriyet Daily News*, Jun 19, 2015, http://www.hurriyetdailynews.com/wikipedia-releases-warning-on-turkeys-censorship-monitoring.aspx?pageID=238&nid=84255.
[6] Jillian C. York, "UK Blocks Access to Wikipedia Entry on Controversial Scorpions Album," *OpenNet Initiative*, Dec 9, 2008, https://opennet.net/blog/2008/12/uk-blocks-access-wikipedia-entry-controversial-scorpions-album.
[7] "Russian media regulator confirms Wikipedia blacklisted," *Russia Beyond the Headlines*, Apr 5, 2013, http://rbth.com/news/2013/04/05/russian_media_regulator_confirms_wikipedia_blacklisted_24706.html. Amar Toor, "Russia banned Wikipedia because it couldn't censor pages," *The Verge*, Aug 27, 2015, http://www.theverge.com/2015/8/27/9210475/russia-wikipedia-ban-censorship.
[8] aacool, "Pakistan Blocks Wikipedia," *Blogcritics*, Mar 31, 2006, http://blogcritics.org/pakistan-blocks-wikipedia/.
[9] "Syrian Youth Break Through Internet Blocks," *IWPR*, http://www.css.ethz.ch/content/specialinterest/gess/cis/center-for-securities-studies/en/services/digital-library/articles/article.html/88422.
[10] "Authorities block access to online encyclopaedia," *Reporters Without Borders / IFEX*, Oct 21, 2005, http://www.ifex.org/china/2005/10/21/authorities_block_access_to_online/.
[11] "New York Times website unblocked, YouTube still inaccessible," *Reporters Without Borders*, Dec 7, 2006, http://archives.rsf.org/print.php3?id_article=20016.
[12] Alice Backer, "Tunisia: Censoring Wikipedia?," *Global Voices*, Nov 27, 2006, https://globalvoices.org/2006/11/26/tunisia-censoring-wikipedia/.
[13] "Uzbekistan Blocks Its Wikipedia," *Sputnik News*, Feb 17, 2013, http://sputniknews.com/world/20120217/171367528.html.

when the site added HTTPS support across all of its various language projects in October 2011.[14] HTTPS makes blocking specific pages on a domain significantly more difficult by preventing censors from seeing exactly which page on a website is being visited, which means censors who want to prevent users from accessing individual Wikipedia articles must choose between blocking the entire site, including inoffensive articles, and not blocking anything at all. The option to access Wikipedia using either the HTTP or HTTPS protocol meant that in some countries where individual articles on the HTTP version of the site had been blocked, the entire site was now available through HTTPS. In China, users suddenly had access to hundreds of previously blocked articles.[15] This lasted for over 18 months, until China blocked the entire HTTPS version of the site in May 2013, forcing users back to the filtered HTTP version.[16] Iran, which in 2013 was found to be blocking more than 1000 individual articles on Persian-language Wikipedia,[17] appears to have left access to the HTTPS version open.

In June 2015, to increase privacy protection and uncensored site access for its users, the Wikimedia Foundation, which hosts Wikipedia, removed the option to access URLs through the HTTP protocol and transitioned fully to HTTPS across all of its sites.[18] While some users lamented the switch, arguing in favor of a "some information is better than none" approach to dealing with censorship, the move was generally perceived by the freedom of expression community as a positive step,[19] the effects of which may already be evident: in August 2015, Russia once again blacklisted Wikipedia over a single cannabis-related article, but the ban was reversed less than 24 hours later.[20]

This report identifies and documents the blocking of Wikipedia content using two complementary data collection and analysis strategies: a client-side system that collected data from the perspective of users around the globe and a server-side tool that analyzed traffic coming in to Wikipedia servers. Collecting and reviewing client-side data allowed us to directly observe censorship; our server-side analysis made use of preexisting data that covered potentially every URL Wikimedia served. Combining these two approaches enabled us to leverage the advantages of each to form a more comprehensive picture of censorship of Wikipedia. Our server-side analysis tracked requests for 1.7 million articles spanning hundreds of languages from November 2011 to late April 2016, as well as

---

[14] Ryan Lane, "Native HTTPS support enabled for all Wikimedia Foundation wikis," *Wikimedia*, Oct 3, 2011, https://blog.wikimedia.org/2011/10/03/native-https-support-enabled-for-all-wikimedia-foundation-wikis/.
[15] "Wikipedia Drops the Ball on China—Not Too Late to Make Amends," *Greatfire*, Jun 3, 2013, https://en.greatfire.org/blog/2013/jun/wikipedia-drops-ball-china-not-too-late-make-amends.
[16] Thomas Fox-Brewster, "Wikipedia Disturbed Over Fresh China Censorship," *Forbes*, May 22, 2015, http://www.forbes.com/sites/thomasbrewster/2015/05/22/wikipedia-disturbed-over-fresh-china-censorship/#295de6885f84.
[17] Nima Nazeri and Collin Anderson, "Citation Filtered: Iran's Censorship of Wikipedia," *Center for Global Communication Studies*, Nov 2013, http://www.global.asc.upenn.edu/fileLibrary/PDFs/CItation_Filtered_Wikipedia_Report_11_5_2013-2.pdf.
[18] Yana Welinder, Victoria Baranetsky, and Brandon Black, "Securing access to Wikimedia sites with HTTPS," *Wikimedia*, Jun 12, 2015, https://blog.wikimedia.org/2015/06/12/securing-wikimedia-sites-with-https/.
[19] Parker Higgins, "Russia's Wikipedia Ban Buckles Under HTTPS Encryption," *Electronic Frontier Foundation*, Aug 28, 2015, https://www.eff.org/deeplinks/2015/08/russias-wikipedia-ban-buckles-under-https-encryption.
[20] Shaun Walker, "Russia briefly bans Wikipedia over page relating to drug use," *The Guardian*, Aug 25, 2015, https://www.theguardian.com/world/2015/aug/25/russia-bans-wikipedia-drug-charas-https.

the general number of requests for each Wikipedia language project from May 2015 through June 2016. Our client-side analysis, which took place primarily in June 2016, covered all of Wikimedia's 292 language projects.

Both client- and server-side methods detected events that we consider likely related to censorship, in addition to a large number of suspicious events that remain unexplained. The blocking of Chinese Wikipedia in China starting in May 2015 was identified in the server-side article data, the server-side project data, and the client-side data. We identified a number of articles that appeared to be censored on Persian Wikipedia prior to the transition to solely HTTPS. Our client-side analysis witnessed transitory but intentional blocking of Yiddish Wikipedia in Thailand, as well as an unconfirmed but highly suspicious inability to access Uzbek Wikipedia from Uzbekistan. This latter event correlated with a highly anomalous decrease in traffic from Uzbekistan to Uzbek Wikipedia apparent in the server-side data. Article analysis uncovered a suspicious decrease in historical traffic to Vietnamese articles related to sex and sexuality. Analysis of project-level data uncovered a number of significant decreases in traffic from various countries that correlated with in-country events. These events ranged from natural disasters to political upheaval and affected access not only to Wikipedia but access to the Internet more broadly.

Methods

While this study has the simply stated goal of analyzing the accessibility of Wikipedia around the world, the methods required are more complex. We broke down the problem into three separate questions: where is Wikipedia blocked, how is Wikipedia blocked, and why is Wikipedia blocked.

To assess where Wikipedia is currently blocked, we used two methods. One looked at the levels of traffic to Wikimedia's servers, and one made requests for the various Wikipedia projects from vantage points around the world. We refer to these two methods respectively as "server-side" and "client-side" analysis throughout the report.

Our server-side data analysis consisted of running an anomaly detection algorithm[21] on the daily number of requests from every country to each of Wikipedia's 292 language projects.[22] This data was available from May 2015 through June 2016, and we were given access to this data on Wikimedia's servers under a non-disclosure agreement. When run against this data, the anomaly detection algorithm output an "anomalousness" score for each day's number of requests, where a negative score meant fewer requests than expected and a positive score meant more requests than expected. The resulting anomalies were then filtered to only the most negative anomalies. Graphs of these anomalous events were generated and then manually reviewed for patterns that might indicate

[21] This algorithm consists mainly of Robust Principal Component Analysis and is described in more detail in Appendix E.
[22] For a list of the projects, see Appendix A.

possible censorship events. For cases in which we were interested in specific countries, we generated graphs regardless of the automatically detected anomalies and manually reviewed these.

Our client-side analysis consisted of performing repeated requests to each of Wikipedia's projects from 41 network vantage points located in 40 countries. These countries were chosen because they made up the entirety of our testing network as of June 2016.[23] From each of our test locations, we requested domains of the pattern "http://(project code).wikipedia.org/wiki/," where "(project code)" is the code given by Wikimedia to each of Wikipedia's various language projects (e.g., "http://en.wikipedia.org/wiki/" for English Wikipedia). It is important to note that because we did not have access at the time of testing to in-country DNS servers, all DNS resolution took place using Google's public DNS servers (8.8.8.8 and 8.8.4.4). This means we were unable to detect any manipulation of requests for Wikipedia that took place only at the DNS level.[24]

For each request we performed, we collected the time it took for the request to complete, the final URL of the response after we followed all redirects, and a screenshot of the resulting page as would be seen by the user. For any request that failed on the initial attempt, we repeated the request until we either received a successful response or it was deemed the domain was likely unavailable from the vantage point. Once all the responses were collected, we reviewed the collected data for any irregularities that might indicate blocking or throttling.

Originally, to answer how Wikipedia might be censored, we intended to use full packet captures of our client tests to identify the precise technological method used to interfere with requests. For example, packet captures could be used to discriminate between IP blocking, injected TCP reset packets, DNS poisoning, injected HTTP redirects, TLS certificate spoofing, or other methods. It is also sometimes possible to identify the use of specific censorship products by looking for distinctive traits they might leave in packet captures.[25] [26] Unfortunately, technical limitations in the deployment of our client network prevented us from collecting these packet captures. Therefore, in witnessed cases of blocking, we could do little but speculate as to the exact technological method of censorship.

Apart from (honest) statements of governments and ISPs, the best way we have to learn about why censors block what they do is to look at historical actions for clues to their motivations. To that end, we used two methods to build context around censorship events that might help us understand motivations. First, we performed traditional research to identify and summarize key themes in the history of censorship in several countries around the world. Second, we attempted to use traffic data to specific Wikipedia articles to locate historical instances of potential censorship with the hope that these historical instances would surface themes and help bolster existing research.

---

[23] The full list of the countries in which our test nodes were located is provided in Appendix B.
[24] Further detail of our client-side collection and its potential drawbacks is provided in Appendix B.
[25] "Behind Blue Coat: Investigations of commercial filtering in Syria and Burma," Nov 9, 2011, *Citizen Lab*, https://citizenlab.org/2011/11/behind-blue-coat/.
[26] Clayton, et al., "Ignoring the Great Firewall of China," Jun 2006, https://www.cl.cam.ac.uk/~rnc1/ignoring.pdf.

Our method of detecting potential censorship of articles using traffic data was fairly intuitive. We started with the hypothesis that if an article has an amount of traffic such that the number of requests per some chosen period of time is rarely zero, and then that article is censored for a sizable portion of its audience, traffic to that article will likely decrease a detectable amount. For example, if an article typically sees around 100 requests per day, and it suddenly drops to 10 requests per day for a week, we can assume something has changed. That change event would then be investigated to identify potential causes. To search for such events, we built an anomaly detection pipeline that could automatically detect significant deviations from the normal pattern of requests.[27] We then detected anomalies in the daily request histories from December 2011 through late April 2016 for approximately 1.7 million articles. Our method of selecting this set of articles was designed to favor articles that we considered more likely to be censored. The final set of 1.7 million articles covered 286 distinct Wikipedia language projects (out of the total 292), 132 of which were represented by more than 10,000 articles. All of the detected anomalies were collected in a database that allowed for easy searching.

It is important to note that daily requests to articles were not broken out by geography. Instead, each data point represented the number of requests in a day for a given article from everywhere on Earth. This meant that we could not definitively attribute any given article anomaly to requests from a particular country. Instead, we could only assume the anomaly was most likely related to the country that constituted the largest share of requests to the article's language project. For example, if we located an anomaly in the request history of an article on Persian Wikipedia, the fact that 83.5% of requests for Persian Wikipedia come from Iran gave us some confidence that the anomaly could be related to Iran. On the other hand, if we located an anomaly in an article on English Wikipedia, we felt that we could not claim the anomaly was related to any single country, as nine countries each contribute more than one percent of the total requests for English Wikipedia.[28] While request data broken out by both article and geography existed, only a small amount of this data was relevant to our analysis, and we therefore opted to use a different data source.[29] This was an unfortunate loss of some of the power of interpretability that we had hoped to achieve with our methodology.

Once we had run all of the article histories through our pipeline, we set about manually reviewing and investigating the anomalies that represented the most severe and longest lasting decreases in request traffic. This manual review phase was both necessary and slow. We found it necessary because large decreases in traffic can be caused by many different processes (national holidays, network outages, articles moving or being redirected, bot activity, etc.), so determining whether or not an anomaly is likely a censorship event is an evidence building process. Unfortunately, the large volume of detected anomalies and the fact that our data analysis process included a good deal of

[27] This is the same as the algorithm used for Wikipedia project-level analysis. For more information about the process and the algorithm (Robust Principal Component Analysis), see Appendix E.
[28] "Wikimedia Traffic Analysis Report - Page Views Per Wikipedia Language - Breakdown," *Wikimedia*, May 2016, https://stats.wikimedia.org/wikimedia/squids/SquidReportPageViewsPerLanguageBreakdown.htm
[29] For a fuller description of this decision, see Appendix E.

manual review meant we were not able to investigate all the significant anomalies individually. A full accounting of our article-level analysis methodology and the issues we encountered while implementing it are provided in Appendix E.

We use three kinds of graphs throughout this report. In the simplest case, we show the number of daily requests for a single article over time. In these graphs, there are vertical colored bars to indicate detected anomalies. Blue vertical bars indicate fewer requests than expected while red bars indicate more requests than expected. The depth of the hue roughly indicates the anomalousness of each anomaly relative to the other anomalies for the same article. Anomaly color bars are also included in project-level graphs where we felt they accurately highlighted important points and are excluded where they hindered interpretation. These project-level graphs do not contain numbers on their vertical axis because the data backing these graphs is only publicly available at a less granular level. Numbers are also omitted on the vertical axis of graphs that depict multiple articles at once, but for a different reason. To account for the varying levels of traffic between articles, the vertical axis depicts the percent change in traffic since the start of the graph period. This effectively normalizes the number of requests across articles, and the axis is indicated as such. Anomaly color bars are omitted on multi-article graphs as they tended to hinder interpretability.

Findings

By Country

Country boundaries are not mirrored in the network topology of the Internet with much fidelity, but the censorship decisions with the broadest impact are often made at the national level, so we believe the state is a useful level of assessment. Below, we have highlighted a number of countries. These countries were chosen because they have either reportedly blocked Wikipedia content at some point in the past or because we have evidence of past or present broader Internet censorship within the country. For each country, we provide a short summary of the history and current state of local Internet filtering. Following that, we include the country-specific results of our data analysis that were the most noteworthy. These results may include tests from client locations, analysis of project-level data, or analysis of article-level data.

China

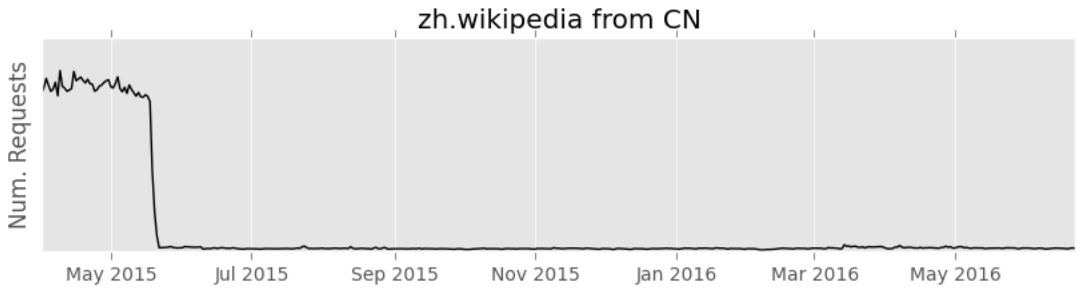
China's Internet filtering apparatus is one of the most pervasive and complex in the world. Freedom House has expressed strong concerns about China's Internet freedoms noting that the country uses a wide variety of techniques—IP blocking, throttling, man-in-the-middle attacks, deep packet inspection, DNS poisoning, keyword filtering, content removal, SMS and instant message filtering, the blocking of VPNs, and full Internet shutdowns in some areas—to block political and sexually explicit content, globally popular social media and publishing platforms, and Google and many of its

services.[30] OpenNet Initiative research conducted from 2004 through 2012 documented extensive, ongoing filtering of political subjects (including the Tiananmen Square protests in 1989; Taiwanese independence; the Uyghur, Tibetan, and Mongolian separatist movements; and criticism of the ruling party); religious subjects (including Falun Gong and the Dalai Lama); international media; human rights groups; pornography; online gambling; social media platforms; and circumvention tools.[31] [32] More recent research has suggested that criticism of the ruling party is largely tolerated while content that has the potential to spur real-world collective action is of primary concern to censors.[33]

Chinese censors have a long and contentious history with Wikipedia. The first Chinese-language Wikipedia project, chinese.wikipedia.org, was launched in May 2001; the first Chinese-language article was published in October 2002, the same month the project moved to zh.wikipedia.org.[34] [35] The project faced its first challenge from censors in June 2004 when it was temporarily blocked during the anniversary of the Tiananmen Square protests.[36] The entire project has been blocked on and off since; article-level filtering of sensitive content was reportedly instituted around 2006.[37] The introduction of an HTTPS version in 2011 temporarily gave users in China full access to the project, including articles blocked on the HTTP site.[38]

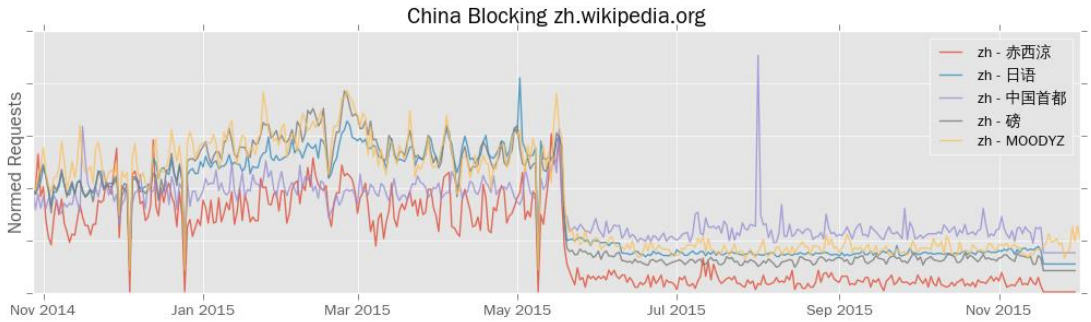
Sophisticated data analysis techniques were not required to identify if and when China has blocked access to Wikipedia. Time series graphs of the number of requests from China to the various Wikipedia projects make it clear when blocking occurred. Below is a graph of the daily number of requests to zh.wikipedia.org from China. A major censorship event is immediately apparent around May 19, 2015:

[30] "China: Freedom on the Net 2015," *Freedom House*, Oct 2015, https://freedomhouse.org/report/freedom-net/2015/china.
[31] "Internet Filtering in China in 2004-2005: A Country Study," *OpenNet Initiative*, 2005, https://opennet.net/studies/china.
[32] "China," *OpenNet Initiative*, Aug 9, 2012, https://opennet.net/research/profiles/china.
[33] Gary King, Jennifer Pan, and Margaret E. Roberts. 2014. "Reverse-engineering censorship in China: Randomized experimentation and participant observation." *Science*, 6199, 345: 1-10, http://gking.harvard.edu/files/gking/files/experiment_0.pdf.
[34] "Chinese Wikipedia," *Wikipedia*, https://en.wikipedia.org/wiki/Chinese_Wikipedia.
[35] Wikipedia offers several Chinese-language projects in addition to zh.wikipedia.org, which is written in Mandarin and automatically translated, based on user preference, into traditional or simplified characters and to incorporate nationally variant vocabulary: Cantonese (https://zh-yue.wikipedia.org), Classical Chinese (https://zh-classical.wikipedia.org), and Min Nan (https://zh-min-nan.wikipedia.org).
[36] Philip P. Pan, "Reference Tool On Web Finds Fans, Censors," *The Washington Post*, Feb 20, 2006, http://www.washingtonpost.com/wp-dyn/content/article/2006/02/19/AR2006021901335.html.
[37] "Wikipedia unblocked in China after year-long ban," *oneindia*, Nov 16, 2006, http://www.oneindia.com/2006/11/16/wikipedia-unblocked-in-china-after-year-long-ban-1163687797.html.
[38] "Wikipedia Drops the Ball on China—Not Too Late to Make Amends," *Greatfire*, Jun 3, 2013, https://en.greatfire.org/blog/2013/jun/wikipedia-drops-ball-china-not-too-late-make-amends.



News reports around the time of this event corroborate that it was indeed caused by intentional government censorship.[39] We analyzed similar graphs for Wikipedia's 291 other language projects and saw no indications of similar anomalies.

While data analysis is not necessary to detect obvious and documented censorship events, our analysis of article-level censorship also picked up this anomaly. As would be expected from this type of censorship, thousands of articles hosted on zh.wikipedia.org saw strong downward anomalies at the same time:



It is important to reiterate that this graph depicts the number of requests to these articles from all geographic locations, not just those requests originating in China. These anomalies are detectable only because a large portion of the worldwide traffic to these Chinese language articles originated in China.

Wikipedia's transition to HTTPS-only delivery occurred in June 2015–almost four weeks after China blocked access to all of zh.wikipedia.org. For that reason, we were unable to analyze the results of the transition to HTTPS-only on the number of requests for Chinese articles.

Using our client network, we were able to confirm that this censorship was ongoing as of late June 2016. We were unable to access zh.wikipedia.org from either of two testing locations in mainland China. While technical limitations in the current deployment of our client network prevent us from

[39] Thomas Fox-Brewster, "Wikipedia Disturbed Over Fresh China Censorship," *Forbes*, May 22, 2015, http://www.forbes.com/sites/thomasbrewster/2015/05/22/wikipedia-disturbed-over-fresh-china-censorship/#295de6885f84.

pinpointing the exact method of censorship, the technological methods of censorship employed by China are extensively documented elsewhere.[40]

We were also able to confirm the result that the zh.wikipedia.org domain was the only Wikipedia project affected by this censorship. Our client machines in both locations were able to successfully and reliably access the other 291 Wikipedia subdomains.[41] In order to check for throughput limitations that may or may not have been intentional ("throttling"), we timed how long it took for a complete response to reach our test clients after sending each request. We refer to this time period as the "round-trip time" ("RTT") throughout. We calculated the mean, median, and max round-trip times to each of the projects from both of our test locations. The results for our tests from China are summarized below:

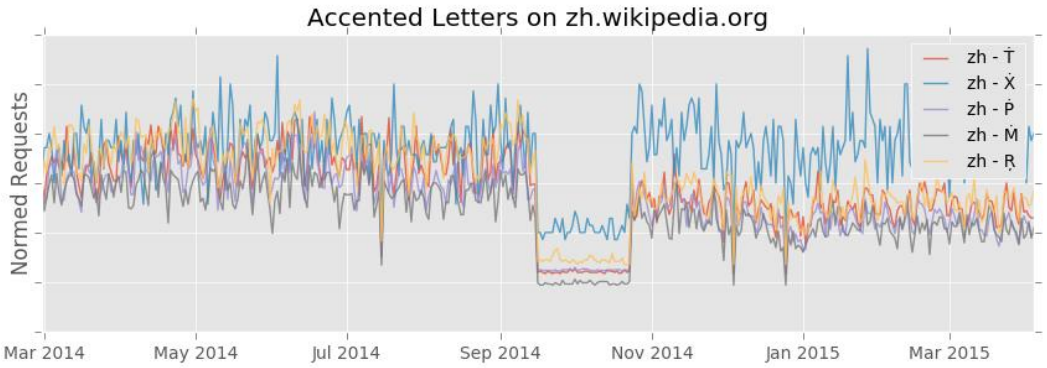
	Median RTT	**Mean RTT**	**Max RTT**
Location 1	550 ms	728 ms	15236 ms to ca.wikipedia.org
Location 2	492 ms	531.8 ms	1492 ms to tw.wikipedia.org

The maximum round-trip time for ca.wikipedia.org is exceptionally long, but subsequent tests were not significantly different from the median.

Our article-level analysis of zh.wikipedia.org found almost 5,500 significant downward anomalies in the number of requests across more than 4,200 articles. A large fraction of these occur around the May 19, 2015 blocking event. While we did not uncover any other article-level events on zh.wikipedia.org that we consider likely censorship, we did encounter events that are both highly anomalous and currently unexplained. For example, articles covering a number of accented letters (Ṗ, Ṁ, Ṫ, Ṛ, Ẋ) saw steep declines in requests beginning September 16, 2014, and all recovered at the same time in mid-October, 2014:

[40] Young Xu, "Deconstructing the Great Firewall of China," *Thousand Eyes*, Mar 8, 2016, https://blog.thousandeyes.com/deconstructing-great-firewall-china/.

[41] One project, Wikipedia in the Nuosu language (ii.wikipedia.org), returned content that was consistent with other Wikipedia projects, but returned a 404 HTTP status code in all our client tests from all locations.



Because these articles contain little content, the number of requests recover overnight, and the article histories show nothing that might explain these changes (such as deleting or renaming the articles), we suspect this behavior might be indicative of either external links to the pages changing or a bot or some other form of programmatic request temporarily suspending activity.

Additionally, our analysis highlighted many anomalous events beginning around August 14, 2013 as well as around August 7, 2015. Articles that were part of these events did not appear thematically related, but traffic drops were significant, and the events were limited to articles in the zh.wikipedia.org domain. While our research did not turn up anything for these dates, we document them here with the hope that they might hold some significance for those more familiar with either Wikipedia's infrastructure or Chinese manipulation of Internet traffic.

While article-level analysis contributed little to the historical context surrounding Chinese Internet censorship, as outlined above, this type of analysis is widely available, and it did serve to bolster our findings from our other methods. Our client tests showed that one Wikipedia domain, zh.wikipedia.org, was completely inaccessible in China, while all other projects were available. Wikimedia's own data on traffic to its projects showed obvious indications of the censorship events reported in the media. While Internet censorship in China is widespread, as of June 2016, Chinese censorship of Wikipedia appears limited to the zh.wikipedia.org domain.

Cuba

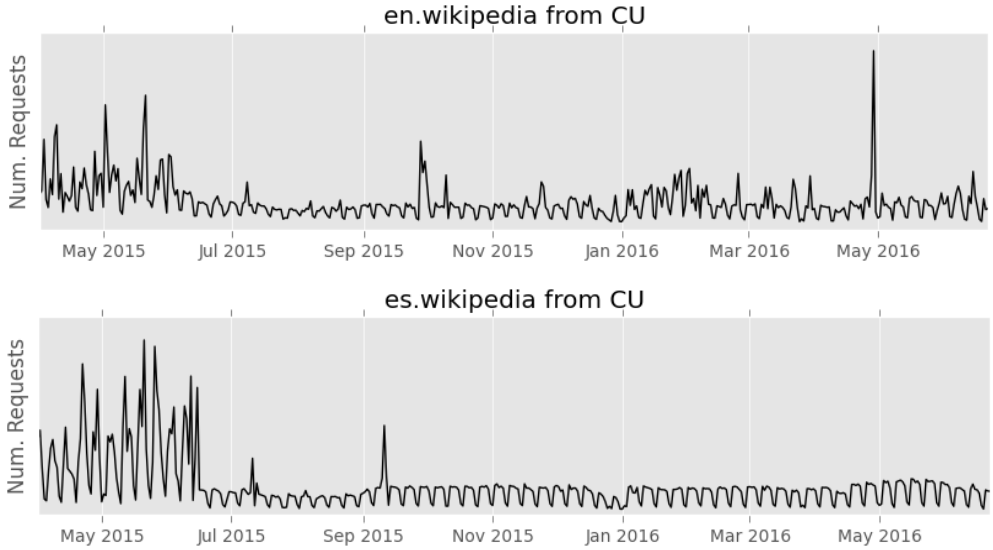
The past three years have seen considerable growth in Cuba's Internet infrastructure, but access is still limited and tightly controlled. The country has two ISPs, both of which are state-owned, and Cuba uses the Avila Link monitoring software to track Internet users and obtain usernames and passwords.[42] Most Cubans are only permitted access to the intranet, which includes a small selection of government-approved websites and services; access to the global public Internet is largely limited to a handful of public WiFi access points and expensive government-run Internet cafes. The Revolutionary Orientation Department (DOR) oversees filtering in the country.[43] Political content,

[42] "Cuba: Freedom on the Net 2015," *Freedom House*, October 2015. https://freedomhouse.org/report/freedom-net/2015/cuba.
[43]"Cuba: Long live freedom (but not for the Internet)!," *Reporters Without Borders*, Mar 12, 2014, http://12mars.rsf.org/2014-en/2014/03/11/cuba-long-live-freedom-but-not-for-the-internet/.

including dissident blogs and news sites, is heavily filtered; common social media platforms such as Facebook and Twitter, VoIP services, and web services such as Yahoo and Hotmail are intermittently blocked.[44]

We did not have a client testing node available in Cuba.

Almost 100% of the requests coming out of Cuba are for either Spanish or English Wikipedia.[45]



Visible in the graphs above is a steep decrease in traffic around the June 12, 2015 HTTPS-only transition. Apart from that, traffic from May 2015 to July 2016 does not show signs that might indicate widespread censorship. Our anomaly detection algorithm did not detect any significant anomalies in the request histories of any other Wikipedia project. While access to the public Internet is restricted, for those with access, we were unable to find any firm evidence that Cuba was censoring any Wikipedia project.

Egypt

Despite offering comparatively free and open access to a wide spectrum of online content, Egypt's Internet environment is still tightly controlled. Political, social, and religious websites are broadly available, but arrests, attacks, self-censorship, and full Internet shutdowns contribute to an atmosphere of repression. Many activists are worried about the draft of a new cybercrime bill introduced in 2015 that would allow the government to heavily increase its censorship role in the name of national security.[46] While this has not yet been enacted, other laws require owners of

[44] Ellery Biddle, "Rationing the Digital: The Policy and Politics of Internet Use in Cuba Today," *Internet Monitor*, Jul 10, 2013, http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2291721.
[45] "Wikimedia Traffic Analysis Report - Wikipedia Page Views Per Country - Breakdown," May 2016, https://stats.wikimedia.org/wikimedia/squids/SquidReportPageViewsPerCountryBreakdown.htm#Cuba
[46] Ragab Saad, "Egypt's Draft Cybercrime Law Undermines Freedom of Expression," *Atlantic Council*, Apr 24, 2015, http://www.atlanticcouncil.org/blogs/menasource/egypt-s-draft-cybercrime-law-undermines-freedom-of-expression.

Internet cafes to track the identities and activities of customers online. VoIP services and encryption tools are also restricted according to Egyptian Telecommunications Laws, but these laws are not widely enforced.

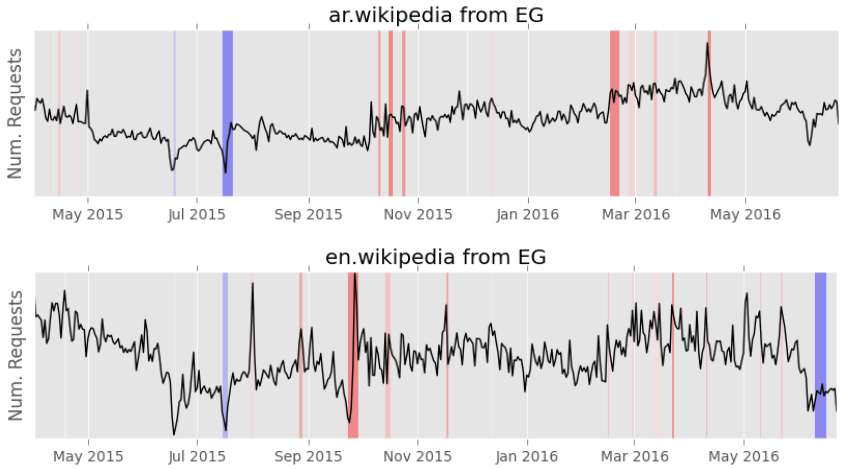
Though it rarely filters online content, the Egyptian government is known for arresting bloggers and journalists critical of the country's current leadership or of Islam. Access to the Internet was most limited during the early 2011 Egyptian revolution protests: for two days both Twitter and Facebook were blocked, and for four days after that, the Internet was down throughout the country. The state's control over the country's telecommunications infrastructure, which is primarily owned by government-run Egypt Telecom, enables the government to slow or completely cut off Internet traffic, mobile messaging, and SMS.[47]

Due to the wide geographic spread of Arabic, historical article anomalies for Arabic Wikipedia are difficult to attribute to Egypt. Noteworthy results from Arabic Wikipedia are presented in Additional Findings, below.

Our client testing node in Egypt was able to successfully and reliably access all Wikipedia projects. Network timing of the responses from each of the projects showed no signs of throttling:

Median RTT	Mean RTT	Max RTT
207 ms	259.1 ms	1792 ms to ceb.wikipedia.org

Analysis of traffic to Arabic and English Wikipedias showed no major anomalies other than during the holidays around the beginning and end of Ramadan:



[47] "Freedom of the Net 2015: Egypt," *Freedom House*, May 2015, https://freedomhouse.org/report/freedom-net/2015/egypt.

While holidays are rarely relevant when discussing the the availability of websites, it is important to note their effect on web traffic, as they can often look similar both statistically and graphically to other types of outage events. A large part of our manual review process was dedicated to successfully ignoring holiday effects.

As of June 2016, we had no evidence that Egypt censored any part of Wikipedia.

Indonesia

Internet censorship in Indonesia is managed by the Ministry of Communication and Information (MCI), which has broad powers to block "negative" content, mostly granted through the Information and Electronic Transactions Law (ITE).[48] MCI maintains a system called Trust Positive[49] which acts as a database cataloguing content that should be censored, but the actual implementation of censorship is left up to the ISPs. As of June 2016, Trust Positive contained approximately 770,000 URLs, about 99.5% of which were categorized as pornographic. Due to the decentralized nature of the censorship infrastructure, some ISPs filter additional URLs while others do not enforce all of the government mandated blocks.[50] For this reason, it is hard to attribute each censored website to the government.

Though most of the content blocked by Indonesian law is pornographic, the relevant statutes are ambiguous, so content related to radicalism, violence, hate speech, fraud, gambling, child violence and pornography, internet security, and intellectual property rights also sees censorship.[51] The pornographic category itself is also very broadly defined. In 2010, the OpenNet Initiative documented evidence of substantial blocking of pornography across different ISPs, but this block also included sites related to women's rights and LGBT websites.[52] Occasionally, LGBT content is specifically targeted by censors, despite being legal in the country.[53] Censors in Indonesia have also appeared willing to censor entire platforms for relatively small amounts of content, at various times blocking all of Netflix, Tumblr, Reddit, and Vimeo, mostly for nudity or sexually explicit content.[54] [55] [56]

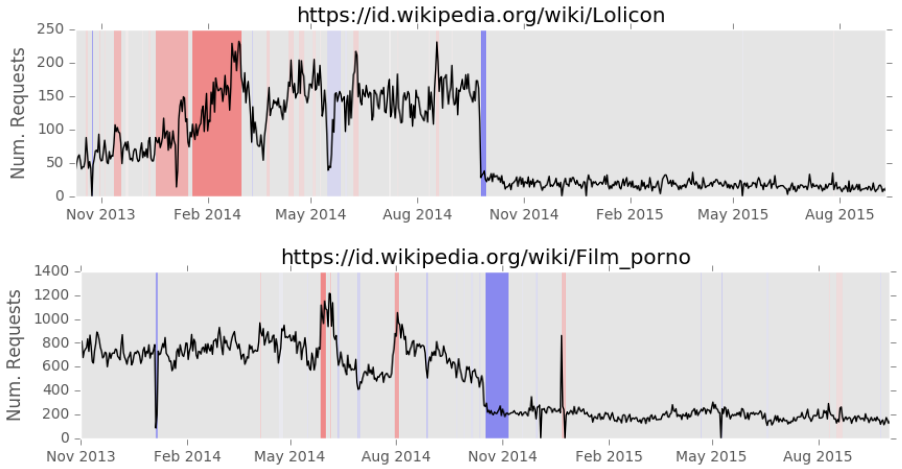
---

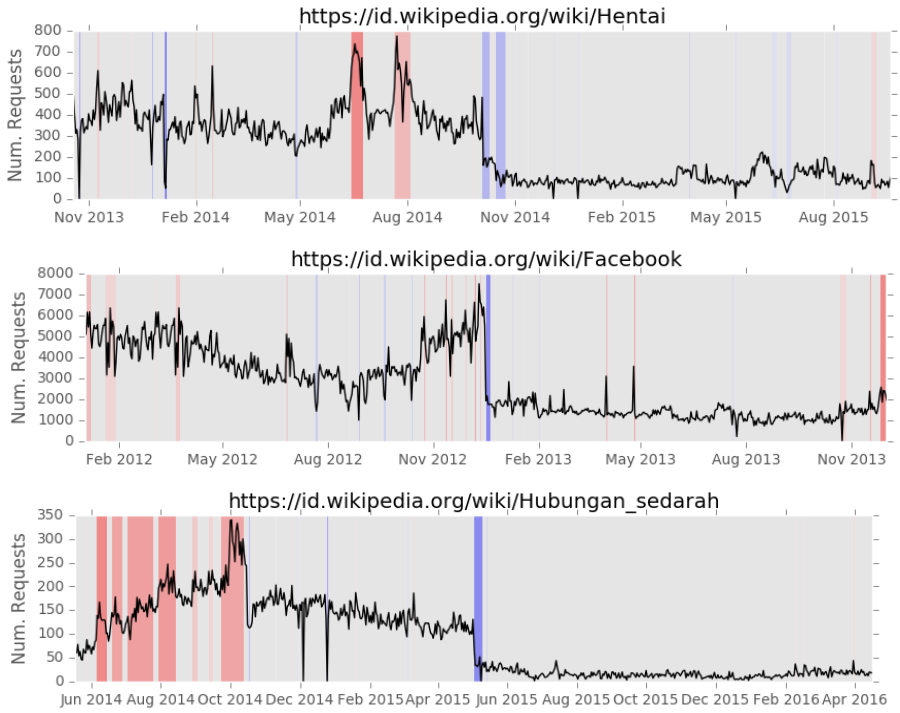
[48] "Freedom on the Net 2015: Indonesia," *Freedom House*, Oct 2015, https://freedomhouse.org/report/freedom-net/2015/indonesia.
[49] See http://trustpositif.kominfo.go.id/.
[50] "Freedom on the Net 2015: Indonesia," *Freedom House*, Oct 2015, https://freedomhouse.org/report/freedom-net/2015/indonesia.
[51] Ibid.
[52] "Indonesia," *OpenNet Initiative*, Aug 9, 2015. https://opennet.net/research/profiles/indonesia.
[53] "Indonesia bans gay emoji and stickers from messaging apps ," *The Guardian*, Feb 11, 2016, https://www.theguardian.com/world/2016/feb/12/indonesia-bans-gay-emoji-and-stickers-from-messaging-apps.
[54] Leo Kelion, "Netflix blocked by Indonesia in censorship row," *BBC*, Jan 28, 2016, http://www.bbc.com/news/technology-35429036.
[55] "Indonesia to ban 477 websites over adult-rated contents," *Xinhua News*, Feb 17, 2016, http://news.xinhuanet.com/english/2016-02/17/c_135106798.htm.
[56] Enricko Lukman, "Amid online porn crackdown, Vimeo, Reddit and Imgur are blocked in Indonesia," *TechInAsia*, May 14, 2014, https://www.techinasia.com/online-porn-crackdown-vimeo-reddit-imgur-blocked-indonesia.

Article-level analysis of Indonesian Wikipedia (id.wikipedia.org) uncovered a large number of anomalies, only some of which could be explained by innocuous causes. While we are not confident enough to claim any of the anomalies we detected were indicative of censorship, we feel a number of anomalies are worth highlighting as suspicious:

Start Date	Indonesian Article	English Article
2012-12-17	Facebook	Facebook
2014-09-25	Lolicon	Lolicon
2014-10-04	Hentai	Hentai
2014-10-17	Film_porno	Pornographic film
2015-05-03	Hubungan_sedarah	Incest

We consider these articles particularly suspicious because most of them are sexual in nature, which, as noted above, is a sensitive topic in Indonesia. None appear to be related to changes to the articles themselves that might otherwise explain significant traffic decreases (such as article deletion or renaming). We do note though that none of the articles show substantial and sustained increases in traffic after the HTTPS-only transition of mid-June, 2015, which we might expect for articles that were censored.

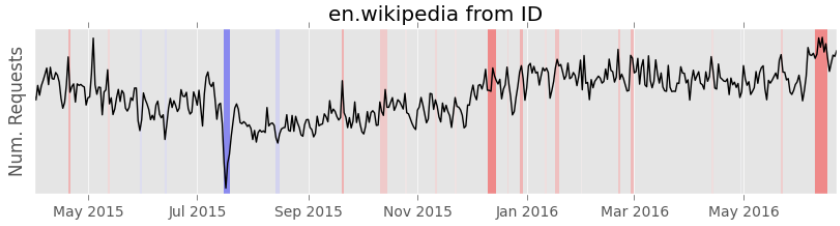




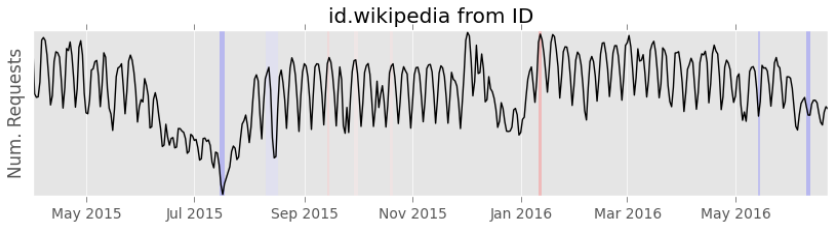
Client side tests from Indonesia returned nothing indicating domain or subdomain blocking. Round-trip times were somewhat slow, but still within normal boundaries. One project, mus.wikipedia.org, took more than four seconds to return, but subsequent requests returned in regular time.

Median RTT	Mean RTT	Max RTT
599 ms	565.4 ms	4588 ms to mus.wikipedia.org

The server-side data on Indonesia did include one significant anomaly that did not appear related to a public holiday. Traffic to Indonesian and English Wikipedias was significantly lower than normal on July 16, 2015. Further research suggests this might have been related to the eruption of two volcanoes, which caused other disturbances throughout the country.[57]



[57] "Indonesia closes three airports as two volcanoes erupt," *Deutsche Welle*, Jul 16, 2015, http://www.dw.com/en/indonesia-closes-three-airports-as-two-volcanoes-erupt/a-18589931.



While it does appear possible that network operators in Indonesia instituted some level of article censorship in the past, our server-side and client-side data analysis did not locate any evidence that censorship of any of Wikipedia's projects was taking place as of June 2016.

Iran

Internet filtering in Iran is implemented by the Commission to Determine the Instances of Criminal Content (CDICC) and broadly overseen by the Supreme Council of Cyberspace; both groups are primarily composed of members appointed by Supreme Leader Ayatollah Ali Khamenei.[58] Content related to the political opposition, human rights (particularly women's rights), minorities, religion, and sex is heavily filtered, as are independent and international media, many major social media platforms, and circumvention tools.[59] [60] President Hassan Rouhani, elected in 2013, promised during his campaign to "ensure that the people of Iran will comfortably be able to access all information globally" and stated that "all human beings have a right" to use social networks.[61] Despite those statements, Facebook, Twitter, and a number of other platforms remain blocked, though Rouhani's administration did resist a CDICC order to block WhatsApp in 2014.[62]

In 2006, then-president Mahmoud Ahmadinejad announced plans to build a national Internet system, in part to improve the country's digital infrastructure and increase speeds, which are currently among the lowest in the world.[63] The project is considerably behind schedule, but is moving forward. One of the project's stated goals is to move the entire country onto a national network, largely disconnected from the greater World Wide Web, to help ensure that Iranian Internet users are accessing "clean" content on domestic Internet hosts.[64] Iran's current filtering technology is already quite centralized: traffic in and out of the country is routed through the

[58] "Iranian Internet Infrastructure and Policy Report," *Small Media*, Apr 2014, https://smallmedia.org.uk/sites/default/files/u8/IIIP_April2014.pdf.
[59] "Iran: Freedom on the Net 2015," *Freedom House*, Oct 2015, https://freedomhouse.org/report/freedom-net/2015/iran.
[60] "Iran," *OpenNet Initiative*, Jun 16, 2009, https://opennet.net/research/profiles/iran.
[61] Saeed Kamali Dehghan, "Hassan Rouhani suggests online freedom for Iran in Jack Dorsey tweet," *The Guardian*, Oct 2, 2013, https://www.theguardian.com/world/iran-blog/2013/oct/02/iran-president-hassan-rouhani-internet-online-censorship.
[62] "Leyla Khodabakhshi," "Rouhani move over WhatsApp ban reveals Iran power struggle," *BBC*, May 8, 2014, http://www.bbc.com/news/world-middle-east-27330745.
[63] "State of the Internet: Q1 2016 Report," *Akamai*, Jun 2016, https://www.akamai.com/us/en/multimedia/documents/state-of-the-internet/akamai-state-of-the-internet-report-q1-2016.pdf.
[64] "Tightening the Net: Internet Security and Censorship in Iran: Part 1: The National Internet Project," *Article19*, Mar 2016, https://www.article19.org/data/files/medialibrary/38315/The-National-Internet-AR-KA-final.pdf.

previously state-owned Telecommunications Infrastructure Company, providing the government with the means to monitor online activities, limit access, throttle speeds, and redirect users attempting to access blocked sites. Authorities also employ keyword filtering, SSL man-in-the-middle attacks, and potentially deep packet inspection to manipulate traffic.[65]

Iran has intermittently blocked access to the HTTPS version of Wikipedia since it was introduced in 2011; the English and Kurdish versions of the site have also seen temporary blocks.[66] In 2013, researchers used proxy servers in Iran to scan every Persian-language Wikipedia URL—approximately 1.7 million in total—and identified nearly 1,000 blocked articles. Just over 400 of these contained political content; the others involved sex, religion, human rights, arts and culture, media and journalists, academia, profanity, drugs, and alcohol. Over half of the blocked articles were biographies of individuals; approximately half of those were biographies of people the government had arrested, detained, or killed. The study concludes that Wikipedia filtering in Iran is in part keyword-based, triggered when users request URLs that match a blacklist of terms; approximately 200 of the articles were filtered on this basis, while the rest were individually blocked.[67] Given this, the transition to HTTPS-only delivery of content in 2015 should have substantially affected the Iranian government's ability to censor Wikipedia articles.

Our article-level analysis indicates that this was indeed the case. We note again that article request histories are not broken out by country; however, Wikimedia's data shows that a large share of the traffic to Persian Wikipedia (fa.wikipedia.org) originates in Iran.[68] Borrowing methodology from another Wikimedia research project,[69] we searched our database of anomalies for articles that saw significantly higher levels of traffic starting around June 12, 2015 (the HTTPS-only transition). We then manually reviewed the resulting articles. This step revealed that many of the articles our algorithm detected saw increased traffic because they were moved or renamed at around the same time as the transition. After removing those articles from our results, we were left with 22 articles that saw increased traffic after the transition that could not be explained by other means.

The set of articles Iran was censoring at the time of the transition was certainly larger than this (as evidenced by the study referenced above), but we do not claim comprehensiveness. We did find that many of the articles identified by our process belonged to the same categories that were most likely to see censorship in the previous research. The set of articles we identified consisted mostly of

---

[65] Simurgh Aryan, Homa Aryan, and J. Alex Halderman, "Internet Censorship in Iran: A First Look," *Proceedings of the 3rd USENIX Workshop on Free and Open Communications on the Internet*, Aug 2013, https://jhalderm.com/pub/papers/iran-foci13.pdf.

[66] "New York Times website unblocked, YouTube still inaccessible," *Reporters Without Borders*, Dec 7, 2006, http://archives.rsf.org/print.php3?id_article=20016.

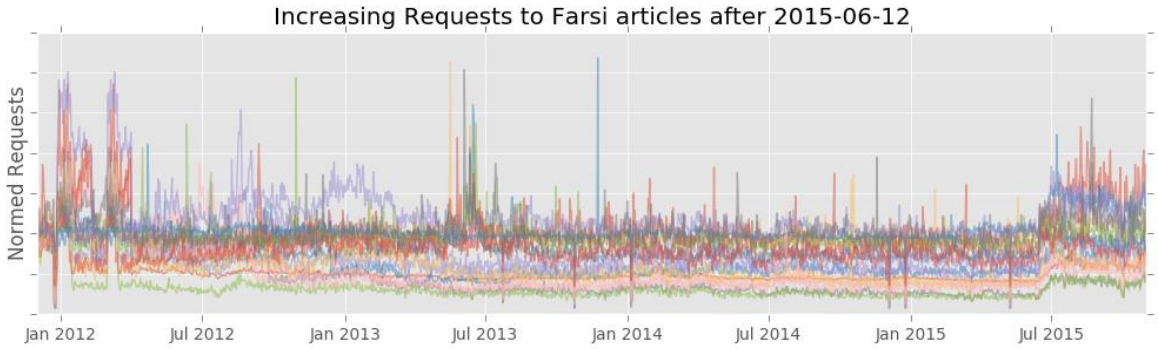
[67] Nima Nazeri and Collin Anderson, "Citation Filtered: Iran's Censorship of Wikipedia," *Center for Global Communication Studies, Annenberg School for Communication (University of Pennsylvania)*, Nov 2013, http://www.global.asc.upenn.edu/fileLibrary/PDFs/CItation_Filtered_Wikipedia_Report_11_5_2013-2.pdf.

[68] "Wikimedia Traffic Analysis Report - Page Views Per Wikipedia Language - Breakdown," May 2016, https://stats.wikimedia.org/wikimedia/squids/SquidReportPageViewsPerLanguageBreakdown.htm#Persian.

[69] "HTTPS Transition and Article Censorship," *Wikimedia*, https://meta.wikimedia.org/wiki/Research:HTTPS_Transition_and_Article_Censorship.

articles related to sex (fifteen articles, e.g., the Persian equivalents of "Sex" and "Cunnilingus"), but also contained political reformers (e.g., the Persian translation of "Mohammad Khatami") and governmental institutions (e.g., the Persian translation of "Army of the Guardians of the Islamic Revolution"). The full list of articles and their English equivalents is included in Appendix C.

Below is a graph of daily traffic to all 22 articles from December 2011 onward:



The uptick in June 2015 is visible, as are two events beginning the end of December 2011 and the end of March 2012 that affect most articles in the set. It is possible Iranian network operators were testing or otherwise adjusting their censorship capabilities around this time, but we were not able to find documented evidence of this.

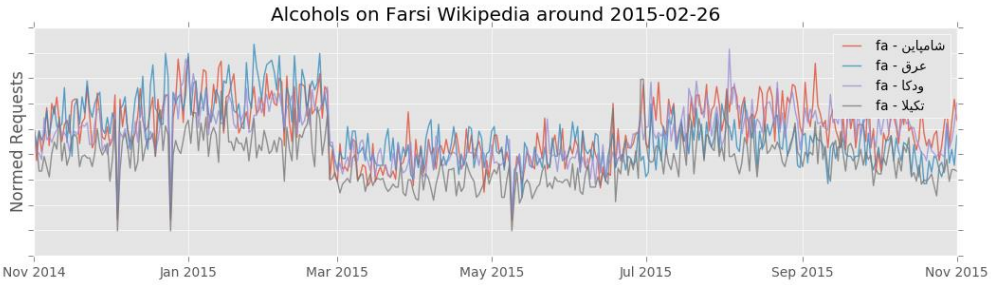
Much of our methodology was designed around locating the beginning of censorship events rather than the end. While this did not produce many positive results, we do believe this method identified the start of a censorship event on Persian Wikipedia. The top four anomalies starting on February 26, 2015 for articles in the fa.wikipedia.org domain are:

Article	Translation
ودکا	Vodka
شامپاین	Champagne
عرق	Sweat
تکیلا	Tequila

Three are clearly identifiable as alcohols, the consumption of which has been illegal in Iran since 1979.[70] The article that translates to "Sweat" is a disambiguation article whose first link is to the article "عرق‌سگی," which translates to "Aragh Sagi." "Aragh Sagi" is a type of alcohol, and "عرق" is a translation of both "sweat" and "distillate." We believe specific censorship of a disambiguation page to be unlikely, and instead suggest that this fact supports the assertion that filtering in Iran is at least

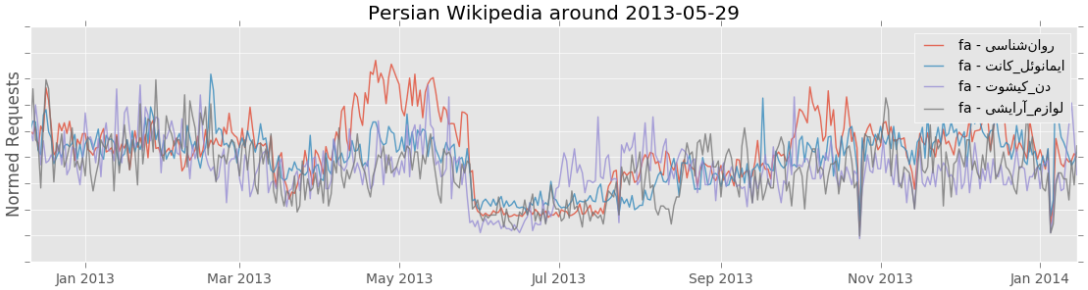
[70] Adam Taylor, "Iran is opening 150 alcoholism treatment centers, even though alcohol is banned," *Washington Post*, Jun 9, 2015, https://www.washingtonpost.com/news/worldviews/wp/2015/06/09/iran-is-opening-150-alcoholism-treatment-centers-even-though-alcohol-is-banned/.

partially keyword based. The graph of daily requests for these four articles around this time period is below:



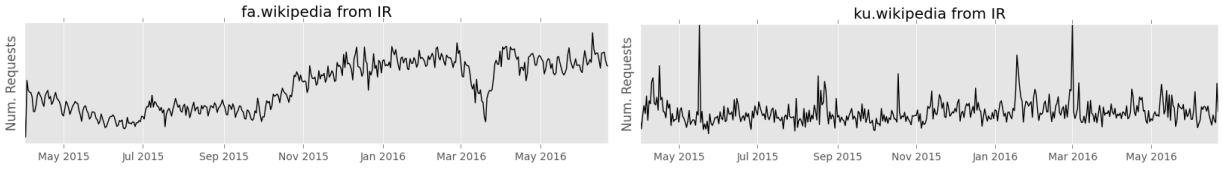
The drop in requests is clearly visible, and while we have not calculated the statistical significance, it appears as if each traffic to each article increases slightly beginning right after the HTTPS-only transition. It is also interesting to note that if this is a censorship event, Iranian censorship officials were actively adding to their lists of blocked content as recently as February 2015, which means these articles were likely censored for only a matter of months.

We also located an event during the spring of 2013 during which more than 20 seemingly unrelated articles saw large falls in traffic (e.g., the following graph depicts traffic to the Persian equivalents of "Psychology," "Immanuel Kant," "Don Quixote," and "Cosmetics"):

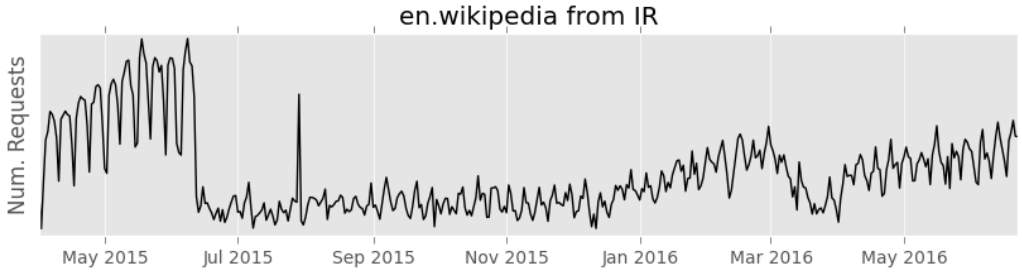


We consider this event unlikely to be censorship because although it happens slightly later, it is similar to many other events across many languages during the spring of 2013 in which numerous unrelated articles saw dramatically decreased traffic before returning to normal levels weeks later. This widespread event is documented in Appendix D.

The number of requests from Iran to Persian and Kurdish Wikipedias—both of which have reportedly been blocked in the past—do not indicate any significant anomalies in the period from May 2015 to July 2016 apart from what is likely decreased traffic due to the holidays around Nowruz (an Iranian holiday celebrating the Iranian New Year) in late March 2016:



Traffic to English Wikipedia from Iran again shows the anomaly that is likely Nowruz plus a rather large drop in traffic around the HTTPS-only transition.



This could have a number of causes, though we believe this decrease is less likely to be related to censorship, as similar decreases in traffic around the time of the transition can be seen in traffic from countries not known to have blocked any part of Wikipedia in the past (e.g., Fiji, outlined in Additional Findings below).

We did not have client testing infrastructure in place in Iran.

Our research on Iran uncovered evidence backing the claims of previous researchers that Iran has blocked Wikipedia articles in the past and that many of those were related to sex or Iranian politics. We further suggest that Wikipedia's transition to HTTPS disabled at least some part of this censorship. While server-side and article analysis indicated that portions of Wikipedia had been censored by Iran in the past, as of late June 2016, evidence of this censorship no longer existed, and at least some articles that had likely seen censorship were receiving increased levels of traffic since Wikipedia's transition to HTTPS.

Kazakhstan

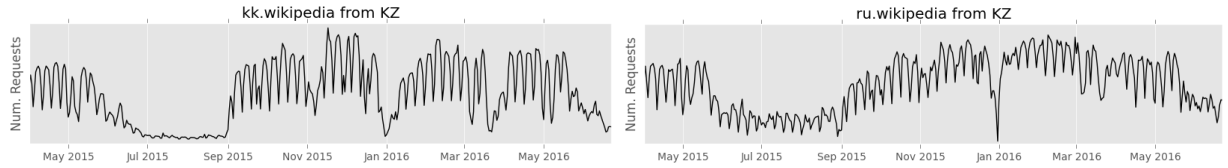
The most heavily censored content in Kazakhstan is that related to religious extremism. Most blocking happens by court order, and throughout all of 2014, the Prosecutor General's Office asked courts to block 703 websites and 198 specific URLs related to the topic. The most significant recent cases of such censorship were related to domestic and international coverage of Kazakhstan's association with ISIS. For example, in the fall of 2014, any web pages containing a series of ISIS videos portraying alleged Kazakh nationals as ISIS soldiers were blocked.[71]

Though the bulk of censorship is dedicated to extremism, popular social media sites have also been targets in the past, though official reasons are rarely given, and ISPs often deny blocking the sites. The blogging platform LiveJournal has been blocked since 2008.[72] Twitter, Facebook, Instagram,

[71] "Kazakhstan: Freedom on the Net 2015," *Freedom House*, Oct 2015, https://freedomhouse.org/report/freedom-net/2015/kazakhstan.
[72] Ibid.

and VKontakte were blocked intermittently for short periods of time in 2014.[73] There have also been several cases of content removal from YouTube, such as a video of ethnic related struggle in South Kazakhstan. Some websites are blocked without any evident court decision, including two major Central Asian news sites, Ca-news (based in Kyrgyzstan) and Fergananews (based in Russia), which are inaccessible for unknown reasons.[74]

Article-level analysis of Kazakh Wikipedia discovered a significant number of anomalies, though further investigation suggested all were associated with the public holidays of either Gregorian New Year or Nowruz (beginning around March 20). Analysis of server-side data revealed much the same thing:



Client-side tests from Kazakhstan were highly inconsistent, with all projects seeing a large number of intermittent errors. These intermittent errors occurred on all tested domains, pointing to an error in the testing node rather than any external issues. Despite this fact, after repeated requests, we were able to successfully receive responses from all Wikipedia projects. The timing of the network requests did not indicate anything out of the ordinary:

Median RTT	Mean RTT	Max RTT
226 ms	239.5 ms	878 ms to iu.wikipedia.org

We were unable to locate any evidence that Wikipedia or any of its projects were being censored in Kazakhstan as of June 2016.

Pakistan

In March 2015, Pakistan's prime minister gave authority over Internet filtering in the country to the Pakistan Telecommunication Authority, skirting existing legislation that vests this power in the Inter-Ministerial Committee for the Evaluation of Web Sites (IMCEW).[75] The change does not yet appear to have affected the country's Internet filtering regime, which is "inconsistent and

[73] "Kazakhstan blocked Facebook, Instagram, twitter and Vkontakte for several hours" [in Russian], *TJournal*, Nov 28, 2014, https://tjournal.ru/p/kazakhstan-total-block.
[74] "Kazakhstan: Freedom on the Net 2015," *Freedom House*, Oct 2015, https://freedomhouse.org/report/freedom-net/2015/kazakhstan.
[75] "Pakistan: Freedom on the Net 2015," *Freedom House*, Oct 2015, https://freedomhouse.org/report/freedom-net/2015/pakistan.

intermittent"[76] but generally targets topic areas that threaten national security or are religiously blasphemous. Access to international news organizations and independent media is generally open, as is access to the websites of human rights organizations, local civil society groups, and Pakistani political parties. Since 2011, all online pornography has been banned, a block that has also affected some sex education and health websites.[77] YouTube has been largely blocked since 2012, when an anti-Islamic video garnered attention throughout the Muslim world.[78] In January 2016, a localized version of YouTube was created that allows the Pakistani government to monitor and take down content deemed inappropriate.[79] In 2013, Citizen Lab researchers documented the use of Netsweeper filters to block political, social, and religious on the network of Pakistan Telecommunication Company Limited, the largest telecommunications company in the country.[80] Facebook and Twitter have received public criticism in the West for limiting access to content at the request of the Pakistani government;[81] in 2014, both platforms republished previously blocked content. Wikipedia is generally accessible, but was blocked for a few hours in 2006 and for several days in 2010.[82] [83]

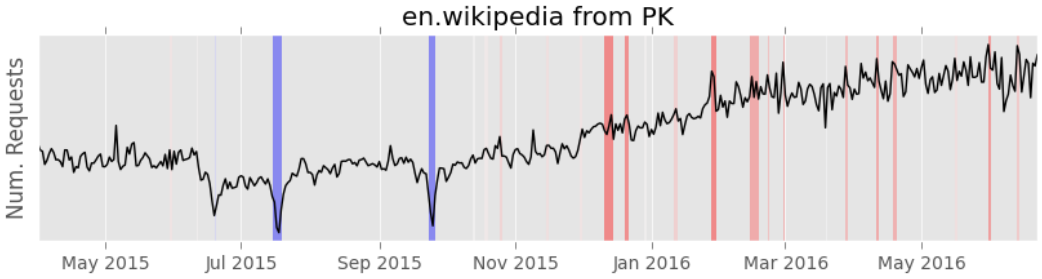
Attributing historical article-level censorship to Pakistan is difficult. As 98% of Wikipedia requests are directed at English Wikipedia,[84] and our current data does not allow us to separate Pakistani requests to English Wikipedia from requests from other countries, we have little-to-no ability to detect Pakistani article censorship.

Our client test node in Pakistan was able to access all Wikipedia projects in a timely fashion:

Median RTT	Mean RTT	Max RTT
281 ms	345.1 ms	1027 ms to am.wikipedia.org

[76] "Pakistan," *OpenNet Initiative*, Aug 6, 2012, https://opennet.net/research/profiles/pakistan.
[77] "Pakistan: Freedom on the Net 2015," *Freedom House*, Oct 2015, https://freedomhouse.org/report/freedom-net/2015/pakistan.
[78] Jon Boone, "Dissenting voices silenced in Pakistan's war of the web," *The Guardian*, Feb 18, 2015, https://www.theguardian.com/world/2015/feb/18/pakistan-war-of-the-web-youtube-facebook-twitter.
[79] Tommy Wilkes, "Pakistan Lifts Ban on YouTube After Launch of Local Version," *Reuters*, Jan 19, 2015, http://www.reuters.com/article/us-pakistan-youtube-idUSKCN0UW1ER.
[80] "O Pakistan, We Stand on Guard for Thee: An Analysis of Canada-based Netsweeper's Role in Pakistan's Censorship Regime," *Citizen Lab*, Jun 20, 2013, https://citizenlab.org/2013/06/o-pakistan/.
[81] "Pakistan - Government Requests Report," *Facebook*, Jan 2014 - Jun 2014, https://govtrequests.facebook.com/country/Pakistan/2014-H1/.
[82] "Websites blocked, PTA tells SC: Blasphemous material," *Dawn*, Mar 14, 2006, http://www.dawn.com/news/183047/websites-blocked-pta-tells-sc-blasphemous-material.
[83] "Pakistan blocks access to YouTube in internet crackdown," *BBC*, May 20, 2010, http://www.bbc.com/news/10130195.
[84] "Wikimedia Traffic Analysis Report - Wikipedia Page Views Per Country - Breakdown," May 2016, https://stats.wikimedia.org/wikimedia/squids/SquidReportPageViewsPerCountryBreakdown.htm#Pakistan.

The only significant downward project-level anomalies to English Wikipedia from Pakistan appear to be the holidays around Ramadan and Eid al-Adha.



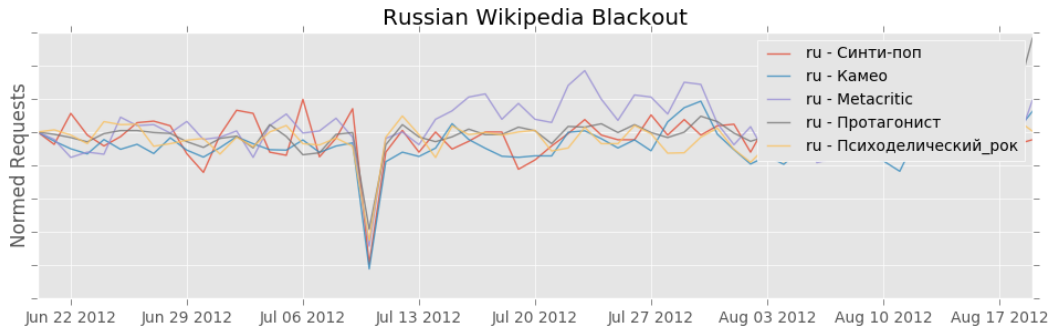
Based on our client tests and Wikipedia data, as of June 2016, we had no firm evidence that any Wikipedia project was being blocked or limited in Pakistan.

Russia

Over the past few years, the Russian government has systematically moved to increase its control over the online information environment, passing new legislation that expands authorities' power to access user data, monitor online activity, and block and take down websites.[85] OpenNet Initiative testing in 2010 found evidence of filtering only of sexually explicit content, but no evidence of political filtering.[86] In the past six years, filtering has grown dramatically and now includes opposition websites, content related to the 2014 conflict in Ukraine and other political protests and events, "extremist" content, and information about drugs and suicide.[87] The federal agency Roskomnadzor, tasked with supervising electronic media in the country, maintains a blacklist of blocked sites; several Wikipedia articles in both Russian and English, most related to drugs or suicide, have reportedly appeared on the list since 2012.[88]

In July 2012, editors of Russian-language Wikipedia shut down the site for 24 hours to protest pending legislation that would increase the government's powers to block online content.[89] This event was represented in our article-level analysis as the most anomalous event we saw for Russian Wikipedia. On July 10, 2012, there were significant decreases in traffic across more than 1,000 articles that quickly disappeared the next day:

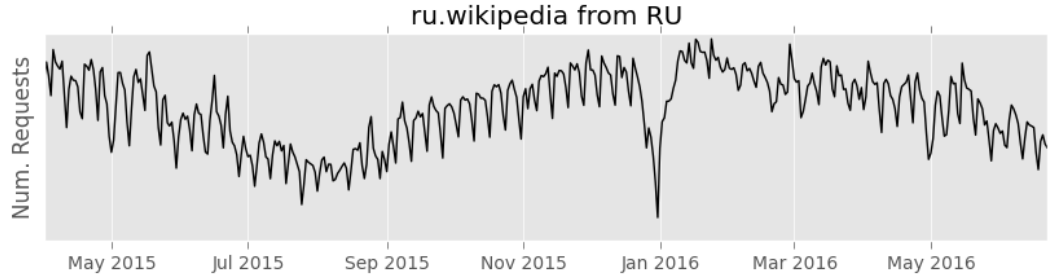
[85] Andrey Tselikov, "The Tightening Web of Russian Internet Regulation," *Berkman Center for Internet & Society (Harvard University)*, Nov 2014, http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2527603.
[86] "Russia," *OpenNet Initiative*, Dec 19, 2010, https://opennet.net/research/profiles/russia.
[87] "Russia: Freedom on the Net 2015," *Freedom House*, Oct 2015, https://freedomhouse.org/report/freedom-net/2015/russia.
[88] "Wikipedia Pages in the Unified Register of Banned Sites" [in Russian], *Wikipedia,* https://ru.wikipedia.org/wiki/Википедия:Страницы_Википедии,_внесённые_в_Единый_реестр_запрещённых_сайтов.
[89] "Russia's Wikipedia shuts down for 24 hours," *ABC News Online*, Jul 10, 2012, http://www.abc.net.au/news/2012-07-11/russias-wikipedia-shuts-down-for-24hrs/4122664.



The fact that about two-thirds of all traffic to Russian Wikipedia originates in Russia[90] supports the conclusion that this event was indeed related to the protest.

In August 2015, access to ru.wikipedia.org was temporarily blocked after Russian Wikipedia did not meet Roskomnadzor's demands to remove an article about a type of cannabis. The site's use of HTTPS meant the internet service providers were unable to block the individual offending page and therefore would have to block all of Russian Wikipedia.[91] The block lasted for several hours before Roskomnadzor announced that the article had been sufficiently edited to meet its guidelines, though Wikipedia editors said the page remained the same.[92] The decrease in traffic that this ban likely caused was not detected by our algorithm on either the article level or the level of Russian Wikipedia as a whole.

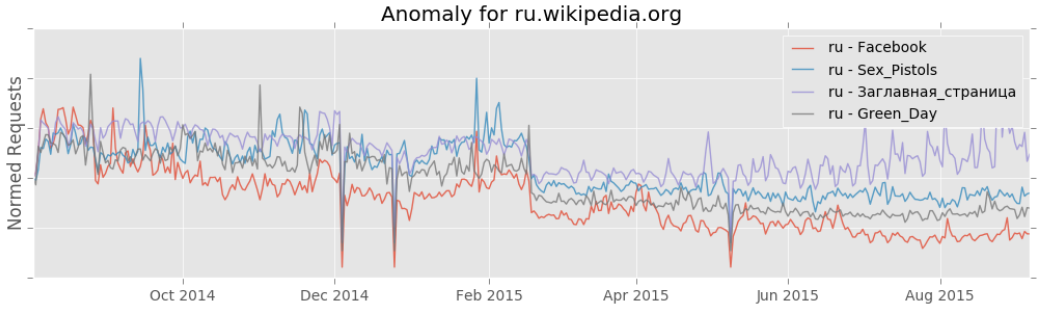
Most of the remaining large anomalous events we detected in article traffic occurred around holidays, most notably the New Year. Across all the Wikipedia projects we looked at, the holiday effect appeared strongest in Russian Wikipedia. Thousands of articles had large decreases starting near the end of 2011, 2012, 2013, 2014, and 2015. The graph of the number of views to all of ru.wikipedia.org from Russia shows one of these strong New Year anomalies:



The remaining anomalies detected by our article-level analysis are unexplained. The most significant of these unexplained events consisted of 261 articles dropping off around February 18, 2015, with

[90] "Wikimedia Traffic Analysis Report - Page Views Per Wikipedia Language - Breakdown," May 2016, https://stats.wikimedia.org/wikimedia/squids/SquidReportPageViewsPerLanguageBreakdown.htm#Russian.
[91] Amar Toor, "Russia banned Wikipedia because it couldn't censor pages," *The Verge*, Aug 27, 2015, http://www.theverge.com/2015/8/27/9210475/russia-wikipedia-ban-censorship.
[92] Shaun Walker, "Russia briefly bans Wikipedia over page relating to drug use," *The Guardian*, Aug 25, 2015, https://www.theguardian.com/world/2015/aug/25/russia-bans-wikipedia-drug-charas-https.

185 articles dropping off on the eighteenth itself. For many of these articles, traffic did not immediately recover. A graph of four of these articles is below:



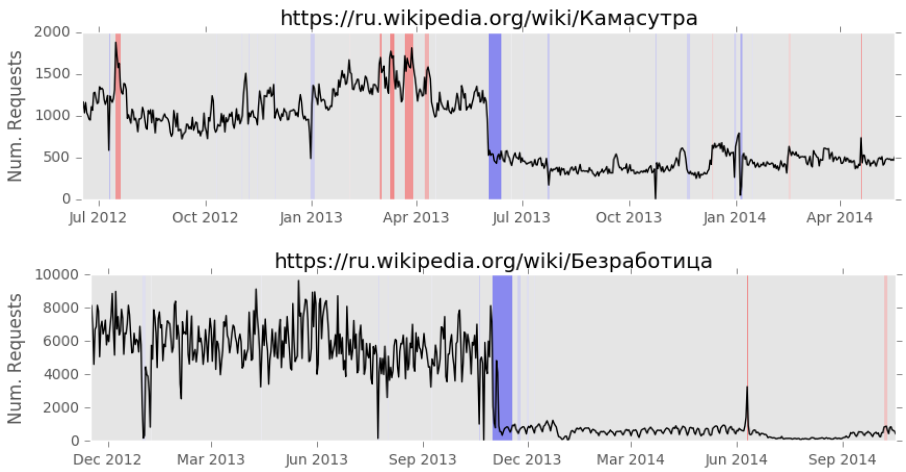
"Заглавная страница" is the Russian equivalent of English Wikipedia's Main Page, and saw an average of more than 800,000 requests per day prior to this event. The decrease in traffic may have been related to the conflict between Russia and Ukraine that was taking place at the time. The average number of monthly requests to Russian Wikipedia from Ukraine for December 2014 and January 2015 was 53,471.5, while the average for February, March, and April of 2015 was 32,105.[93] We were unable to find any other evidence to support this hypothesis.

Russian Wikipedia also contained a relatively large number of anomalies that were limited in scope to single articles. Investigation of many of these cases revealed that in most circumstances, the articles in question were deleted or moved (e.g. "Нагота" ["Nudity"] on July 9, 2014, "Косово" ["Kosovo"] on January 8, 2013, and "Массовое_убийство" ["Mass Murder"] on February 20, 2014). These were picked up by the anomaly detection algorithm, as they often had a significant amount of traffic prior to deletion. After manually removing from analysis those articles that had plausible explanations for traffic drops, we were still left with a number of articles with unexplained significant traffic drops:

Start Date	Article	Translation
2013-05-11	Уэйко	Waco
2013-06-02	Камасутра	Kamasutra
2013-07-30	Кроманьонцы	Cro-Magnon
2013-11-01	Безработица	Unemployment
2013-11-14	Анис_обыкновенный	Anise

Sample graphs of anomalies detected in "Unemployment" and "Kamasutra" are below:

[93] "Wikimedia Traffic Analysis Report - Wikipedia Page Views Per Country - Breakdown," Dec 2015 - April 2016, *Wikimedia*, https://stats.wikimedia.org/archive/squid_reports/2015-02/SquidReportPageViewsPerCountryBreakdownHuge.htm#Ukraine.



From our client test node in Russia, we were able to access all Wikipedia project subdomains successfully and reliably. Network request round trip times were the fastest of all we tested:

Median RTT	Mean RTT	Max RTT
96 ms	118.8 ms	688 ms to bxr.wikipedia.org

While Russia has actively censored portions of the Internet, and as of June 2016, that censorship appeared to be growing, we found no evidence that Russia was interfering with traffic to Wikipedia at either the article or project level.

Saudi Arabia

In 2014, Reporters without Borders ranked the Kingdom of Saudi Arabia 164th out of 180 countries in terms of press freedom, emphasizing that the Kingdom is "relentless in its censorship of the Saudi media and the Internet."[94] All international Internet traffic is routed through two national providers, Integrated Telecom Company and Bayanat al-Oula for Network Services, giving the government the ability to review and filter requests.[95] The Communications and Information Technology Commission oversees Internet filtering in the country, and the list of content blocked in the country is long. First, Saudi Arabia uses commercially available software (SmartFilter[96]) to locate URLs related to pornography, gambling and drugs, which it then blocks. They also maintain a local list of

[94] "World Press Freedom Index 2014," *Reporters Without Borders*, Jan 31, 2014, https://rsf.org/sites/default/files/index2014_en.pdf.
[95] "Saudi Arabia: Freedom on the Net 2015," *Freedom House*, Oct 2015, https://freedomhouse.org/report/freedom-net/2015/saudi-arabia.
[96] Jakub Dalek, et al., "A Method for Identifying and Confirming the Use of URL Filtering Products for Censorship," *Sigcomm ICM*, Oct 2013, http://conferences.sigcomm.org/imc/2013/papers/imc112s-dalekA.pdf.

URLs separate from this categorization mechanism.[97] This list reportedly contains a broader set of content, including content related to violent extremism, criticism of Gulf royal families, political opposition, censorship circumvention tools, P2P file sharing tools, LGBT issues, human rights organizations, religious scholars (especially those related to the minority Shi'a faith), mirror sites, and unlicensed online publications.[98] [99] It is unclear how willing Saudi authorities are to block entire sites over single pieces of content. In 2012, the government threatened to block YouTube if a controversial video was not taken down, but the blocking did not occur because YouTube removed the video in question.[100]

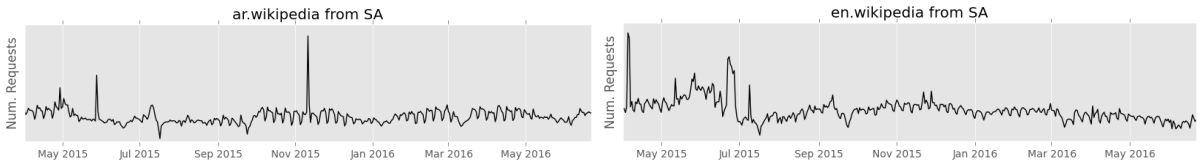
Internet restrictions in Saudi Arabia are not limited to content filtering; a 2009 law led to the installation of hidden cameras in all web cafes to track users, and self-censorship among online writers is widespread.[101] The government regularly arrests those who use social media to document human rights abuses, express political opinions critical of the ruling family, or criticize the official religion; those who are convicted are sentenced to jail time and, in at least one case, corporal punishment.[102]

In 2006, Saudi Internet users started reporting the censorship of a number of Wikipedia pages in both English and Arabic, mostly related to sexual content.[103] When the blocking occurred, some Saudi citizens felt that some of the pages were unfairly blocked and contained "beneficial" content.[104]

Arabic and English Wikipedia together account for more than 95% of the requests from Saudi Arabia,[105] and our analysis did not show the type of traffic anomaly that would be indicative of domain blocking over the period from May 2015 to July 2016.

---

[97] "General Information on Filtering Service," *Saudi CITC*, http://www.internet.sa/en/general-information-on-filtering-service.
[98] "Saudi Arabia: Freedom on the Net 2015," *Freedom House*, Oct 2015, https://freedomhouse.org/report/freedom-net/2015/saudi-arabia.
[99] "Internet Filtering in Saudi Arabia," *OpenNet Initiative*, Aug 6, 2009, https://opennet.net/research/profiles/saudi-arabia.
[100] "YouTube blocks 'Innocence of Muslims' in Saudi Arabia," *Al Arabiya News*, Sep 19, 2012, http://english.alarabiya.net/articles/2012/09/19/238987.html.
[101] "Internet Filtering in Saudi Arabia," *OpenNet Initiative*, Aug 6, 2009, https://opennet.net/research/profiles/saudi-arabia.
[102] Ben Beaumont, "7 Ways Saudi Arabia is Silencing People Online," *Amnesty International*, Apr 9, 2015. https://www.amnesty.org/en/latest/campaigns/2015/04/7-ways-saudi-arabia-is-silencing-people-online/.
[103] "List of Wikipedia articles censored in Saudi Arabia," *Wikipedia*, https://en.wikipedia.org/wiki/Wikipedia:List_of_Wikipedia_articles_censored_in_Saudi_Arabia.
[104] Hassana'a Mokhtar, "What is Wrong With Wikpedia," *Arab News*, Jul 19, 2006, https://web.archive.org/web/20110807060237/http://archive.arabnews.com/?page=1&section=0&article=85616&d=19&m=7&y=2006.
[105] "Wikimedia Traffic Analysis Report - Wikipedia Page Views Per Country - Breakdown," May 2016, https://stats.wikimedia.org/wikimedia/squids/SquidReportPageViewsPerCountryBreakdown.htm#Saudi_Arabia.



We were able to access all Wikipedia subdomains from our client test point in Saudi Arabia, and all round-trip times were within normal ranges:

Median RTT	Mean RTT	Max RTT
283.5 ms	308.6 ms	1142 ms to am.wikipedia.org

Our article-level analysis is not segmented by country, and while the largest share of requests to Arabic Wikipedia come from Saudi Arabia, that share is only approximately one-fifth.[106] If we were to locate likely censorship events in Arabic Wikipedia, it would be impossible without additional data to definitively attribute that censorship to Saudi Arabia. Given the results of our client and server data analysis, as of June 2016 we had no firm evidence that Saudi Arabia was censoring any Wikipedia domain or subdomain.

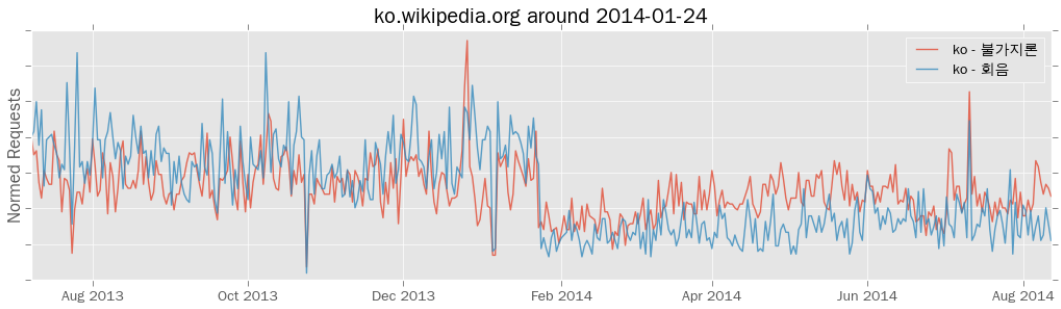
South Korea

South Korea's Internet filtering regime is largely focused on its relations with North Korea and on sexually explicit content. The majority of banned websites are North Korean news organizations or sites run by North Korean "sympathizers," but pornography and LGBT websites are also widely banned.[107] The National Security Act in Cyberspace prohibits, among other things, "sympathizing" with North Korea online; more than 100 people were convicted of this crime between 2012 and 2014.[108] South Korea's constitution states that "neither speech nor the press may violate the honor or rights of other persons nor undermine public morale or social ethics."[109] These restrictions have been used to justify the censoring of attacks against politicians, sites connected to North Korea, and pornography sites. The government's decision to ban online gaming for six hours each day for citizens younger than sixteen also loosely falls under this guideline.[110]

[106] "Wikimedia Traffic Analysis Report - Page Views Per Wikipedia Language - Breakdown," May 2016, https://stats.wikimedia.org/wikimedia/squids/SquidReportPageViewsPerLanguageBreakdown.htm#Arabic.
[107] "South Korea," *OpenNet Initiative*, Aug 6, 2012. https://opennet.net/research/profiles/south-korea.
[108] Ibid.
[109] "Freedom on the Net 2015: South Korea," *Freedom House*, Oct 2015, https://freedomhouse.org/report/freedom-net/2015/south-korea.
[110] "Why South Korea is really an internet dinsosaur," *The Economist*, Feb 10, 2014, http://www.economist.com/blogs/economist-explains/2014/02/economist-explains-3.

The Korean Communications Standards Commission (KCSC) is in charge of regulating the Internet, but in 2014 the Public Prosecutor's office set up an investigative unit charged with monitoring online slander and rumors.[111] South Korea has a history of defamation cases involving the Internet; in 2012 a National Intelligence Service (NIS) agent removed Twitter accounts that were critical of President Park Geun-hye, who was running for reelection at the time.[112] Just two years later, Han Sun-Kyo, a conservative, attempted to pass a law that would prevent "rumor mongering" in the wake of the capsizing of the Sewol ferry, which left over 300 people dead.[113] Harsh punishments for defamation exist in South Korea; online defamation is penalized severely, with fines reaching $45,000 USD at times.[114]

The article-level analysis we conducted revealed some anomalies that could not be attributed to changes to the articles themselves. There were only two anomalies that occurred at approximately the same time: "회음" ("Perineum") and "불가지론" ("Agnosticism"):



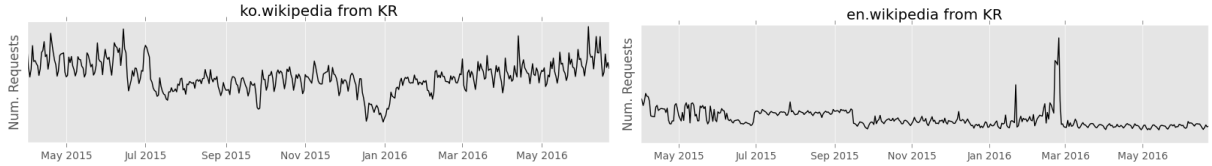
While it is interesting that traffic to both articles dropped significantly on the same day, the fact that this anomaly was limited to these two articles and that they are not closely related thematically makes us doubt that this was a censorship event. There were other anomalous events for single articles throughout our analysis, the most significant of which were "엔진오일" ("Engine Oil") starting on February 19, 2014 and "아가" ("Song of songs") starting on May 17, 2014.

We were able to access all Wikipedia project subdomains from our test location in South Korea with no problems. Response times for each of the domains were within typical ranges:

[111] "Freedom on the Net 2015: South Korea," *Freedom House*, Oct 2015, https://freedomhouse.org/report/freedom-net/2015/south-korea.
[112] Ibid.
[113] Ibid.
[114] Ibid.

Median RTT	Mean RTT	Max RTT
301 ms	316.6 ms	1286 ms to als.wikipedia.org

The history of requests from South Korea to both Korean and English Wikipedias over the period of analysis appear regular with no signs of outages:



As of June 2016, we were unable to find any strong evidence that South Korea has censored or was censoring any of Wikipedia's articles or projects.

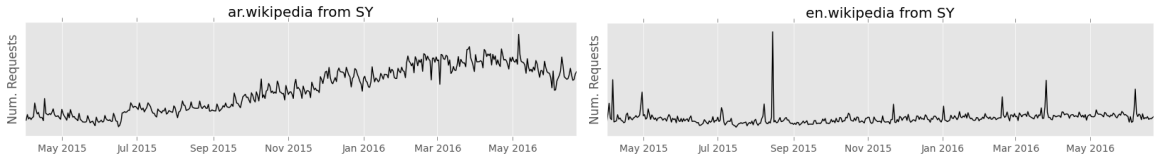
Syria

Syrian netizens experience extensive censorship online around politics, minorities, human rights, and foreign affairs. Examples of censored content include the London-based news outlets *Al-Quds al-Arabi* and *Asharq al-Awsat*, many Lebanese online newspapers, websites campaigning to end Syrian influence in Lebanon, WhatsApp, the Muslim Brotherhood, websites that advocate for the Kurdish minority, and the entire Israeli top-level domain ".il." Websites related to human rights awareness such as the Violations Documentation Center are also blocked.[115] According to the Wall Street Journal in 2012, out of 2,500 attempts to visit Facebook, two-fifths were permitted and three-fifths were blocked.[116] Censorship also extends to mobile communication: Bloomberg reported in 2012 that a special government unit known as Branch 225 had ordered Syrian mobile providers to block text messages containing words like "revolution" or "demonstration."[117] The fact that both YouTube and some pages on Facebook remain accessible make activists suspect that the current regime is trying to track citizens' online activities. Other social media applications like the VoIP service Skype suffer from disruptions either due to low speeds or intermittent blocking by the authorities. Over the past decade authorities have detained hundreds of Internet users, including several well-known bloggers and citizen journalists.[118] Wikipedia in Arabic was reportedly blocked from April 2008 until February 2009, but other languages remained accessible.[119]

[115] "Syria: "Freedom of the Net," *Freedom House*, May 2015, https://freedomhouse.org/report/freedom-net/2015/syria.
[116] Jennifer Valentino-Devries, Paul Sonne, and Nour Malas, "U.S. Firm Acknowledges Syria Uses Its Gear to Block Web," *Wall Street Journal*, Oct 29, 2011, http://on.wsj.com/t6YI3W.
[117] Ben Elgin and Vernon Silver, "Syria Disrupts Text Messages of Protesters With Dublin-Made Equipment," *BloombergBusiness*, Feb 14, 2012, http://bloom.bg/1i0TOEU.
[118] "Syria: "Freedom of the Net," *Freedom House*, May 2015, https://freedomhouse.org/report/freedom-net/2015/syria.
[119] "Syrian Youth Break Through Internet Blocks," *IWPR*, http://www.css.ethz.ch/content/specialinterest/gess/cis/center-for-securities-studies/en/services/digital-library/articles/article.html/88422.

Using our methodology and the data available, article-level censorship would be difficult to attribute to Syria, as Arabic Wikipedia is accessed heavily from many countries. We did not have a client test node in Syria.

Our analysis of server-side data detected no significant anomalies in traffic from Syria to any Wikipedia project. Nevertheless, we conducted a manual review of Arabic and English Wikipedia because they are the most popular Wikipedia projects in Syria, together accounting for approximately 98% of traffic.[120] The number of requests to these Wikipedia projects show no significant anomalies between May 2015 and July 2016:



While censorship of the Internet is known to be widespread in Syria, and censorship of Wikipedia specifically has occurred in the past, the lack of both data and access made establishing the June 2016 state of Wikipedia in Syria particularly difficult.

Thailand

Censored content in Thailand is similar to that of other countries: pornography, gambling, and censorship circumvention tools are all extensively blocked,[121] but the censorship extends to content that is specifically sensitive in the context of Thailand. As there have been a number of coups in Thailand in recent years, political opposition and activism content is strongly suppressed, as are some foreign news outlets, some domestic news outlets, human rights content, select academic websites, and Facebook and YouTube pages that relate to coups.[122]

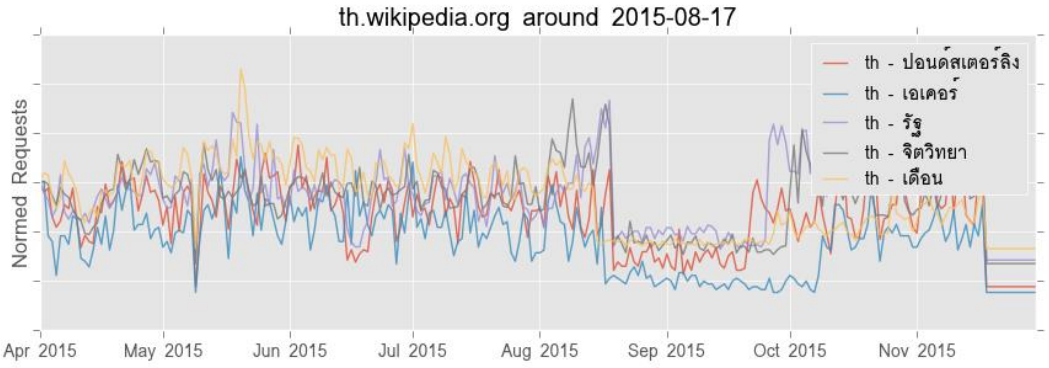
Lèse-majesté, the insult or defamation of royalty, is a serious crime in Thailand, and has lead to a number of censorship incidents. The law against lèse-majesté has been used to prosecute those who have posted social media updates, news articles, audio and video content, and poetry deemed offensive, as well as at least one Internet user who sent an email containing links to *lèse-majesté* content. In 2015, prison sentences for violating the prohibition on *lèse-majesté* reached a record high of 60 years.[123] In 2008, the Wikipedia article for Bhumibol Adulyadej, the King of Thailand, was

[120] "Wikimedia Traffic Analysis Report - Wikipedia Page Views Per Country - Breakdown," May 2016, https://stats.wikimedia.org/wikimedia/squids/SquidReportPageViewsPerCountryBreakdown.htm#Syria
[121] "Internet Filtering in Thailand," *OpenNet Initiative*, Aug 7, 2012, https://opennet.net/research/profiles/thailand.
[122] "Thailand: Freedom on the Net 2015," *Freedom House*, Oct 2015, https://freedomhouse.org/report/freedom-net/2015/thailand.
[123] "Thailand: Freedom on the Net 2015," *Freedom House*, Oct 2015, https://freedomhouse.org/report/freedom-net/2015/thailand.

reportedly blocked.[124] No official reason was given, but it was possibly related to the lèse-majesté law.

The military junta that took power during the 2014 coup intensified controls over the Internet, instituting new filtering and surveillance and arresting activists and others.[125] During the coup, the government ordered ISPs to block Facebook in an effort to prevent activists from protesting.[126] Those who criticized the coup online were detained and forced to promise silence and turn over their social media passwords in exchange for their release; the government collected 400 passwords this way in 2014.[127] In August 2015, the government announced plans to implement a "Great Firewall" that would have directed all Internet traffic through a single point, but it abandoned the plan in October 2015.[128]

Article-level analysis of Thai Wikipedia did not reveal anything we would consider likely censorship events, though there was an anomalous event that we were not able to explain. Starting around August 17, 2015, a number of thematically unrelated articles saw significant decreases in traffic that each lasted for about a month before returning to previous levels. This anomaly took place during the period of time for which we had request data broken out by both article and geography, and we were able to confirm that this anomaly was present in requests originating in Thailand. A graph depicting the event is below:



[124] "Strange Thai Internet Censorship," *YouTube*, Nov 16, 2008, https://www.youtube.com/watch?v=6ToQ4zgQZ_E.
[125] "Information Controls During Thailand's 2014 Coup," *Citizen Lab*, Jul 9, 2014, https://citizenlab.org/2014/07/information-controls-thailand-2014-coup/.
[126] Ibid.
[127] "Thailand: Freedom on the Net 2015," *Freedom House*, Oct 2015, https://freedomhouse.org/report/freedom-net/2015/thailand.
[128] Amy Sawitta Lefevre, "Thailand scraps unpopular internet 'Great Firewall' plan," *Reuters*, Oct 15, 2015, http://www.reuters.com/article/us-thailand-internet-idUSKCN0S916I20151015.

Article	Translation
ปอนด์สเตอร์ลิง เอเคอร์ รัฐ จิตวิทยา เดือน	Pound Sterling Acre State Psychology Month

While we are unsure of the cause of the event, it is unlikely to be article-level censorship as this event took place after the June 2015 transition to HTTPS-only content delivery.

Our client test node in Thailand was able to successfully access all of the Wikipedia domains at least once, but we did witness a single censorship event. On June 25, 2016 one test to Yiddish Wikipedia (http://yi.wikipedia.org/wiki) was redirected to http://203.113.26.210/?v0.42p3, which returned a

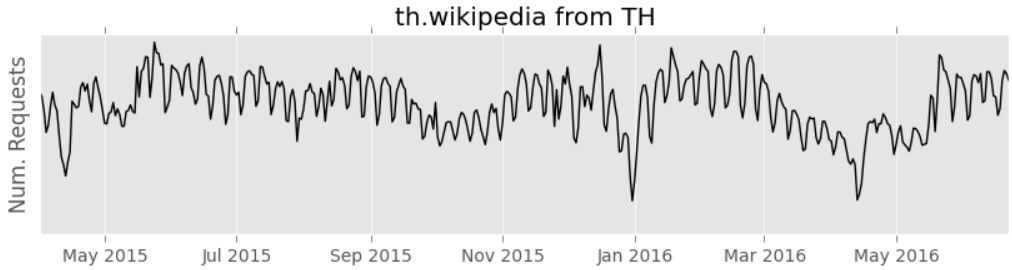
page that contained only this image:
The text in the image translates roughly to: "This website contains information that is not appropriate. Suspended by the Ministry of Information and Communication Technology." Ten additional requests to Yiddish Wikipedia over the course of a week and 1,818 requests in total to the other Wikipedias did not result in receiving this page again. We were unable to find evidence of prior censorship of Yiddish material in Thailand or much evidence of a Yiddish-speaking population. It is possible that this censorship event might have been the result of a temporary misconfiguration, which would not be unprecedented.[129]

Round-trip times to the various Wikipedias were the slowest of all our testing locations, but they were still within acceptable ranges.

[129] Amy Qin, "Chinese Web Outage Blamed on Censorship Glitch," *Sinosphere - NYTimes*, Jan 22, 2014, http://sinosphere.blogs.nytimes.com/2014/01/22/chinese-web-outage-blamed-on-censorship-glitch/.

Median RTT	Mean RTT	Max RTT
803 ms	911.1 ms	3806 ms to ee.wikipedia.org

Analysis of server-side data of the number of requests to Thai Wikipedia from Thailand shows no significant anomalies apart from the holidays around Gregorian New Year and Thai New Year (Songkran, beginning April 13 in both 2015 and 2016). Traffic to Yiddish Wikipedia from Thailand is too low volume to be able to identify any anomalously low periods.



While it is unlikely we identified article-level censorship in Thailand, we did confirm that as of June 2016, Thailand was at least intermittently interfering with the regular functioning of Wikipedia.

Turkey

Internet penetration and usage in Turkey has been rapidly increasing over the last decade with 2014 marking the first year more than half the Turkish population could be considered Internet users.[130] The dramatic increase in Internet usage has seen a concomitant increase in the Turkish government's efforts at controlling access to information on the Internet. Before 2007, Internet censorship in Turkey was sporadic and limited,[131] but the passage of Internet Law No. 5651 in May 2007 was the first big step toward systematizing Turkey's blocking regime and grounding it in a legal framework.[132] Among other things, Law 5651 outlined eight categories of content that were to be subject to blocking, required all Internet hosting and access providers in Turkey to obtain a license from the government, and granted an organization called the Presidency of Telecommunication and Communication (TIB) the authority to block any website it deemed in violation of the law.[133] Since the large, anti-government protests in Gezi Park in June of 2013 in which social media played a large role, Law 5651 has been amended numerous times to relax requirements for judicial review, broaden

[130] "Percentage of Individuals using the Internet," *ITU*, 2015, http://www.itu.int/en/ITU-D/Statistics/Documents/statistics/2016/Individuals_Internet_2000-2015.xls.
[131] Mustafa Akgül and Melih Kırlıdoğ, "Internet censorship in Turkey," *Internet Policy Review*, Jun 3, 2015, http://policyreview.info/articles/analysis/internet-censorship-turkey.
[132] "Turkey," *OpenNet Initiative*, Dec 18, 2012, https://opennet.net/research/profiles/turkey.
[133] Mustafa Akgül and Melih Kırlıdoğ, "Internet censorship in Turkey," *Internet Policy Review*, Jun 3, 2015, http://policyreview.info/articles/analysis/internet-censorship-turkey.

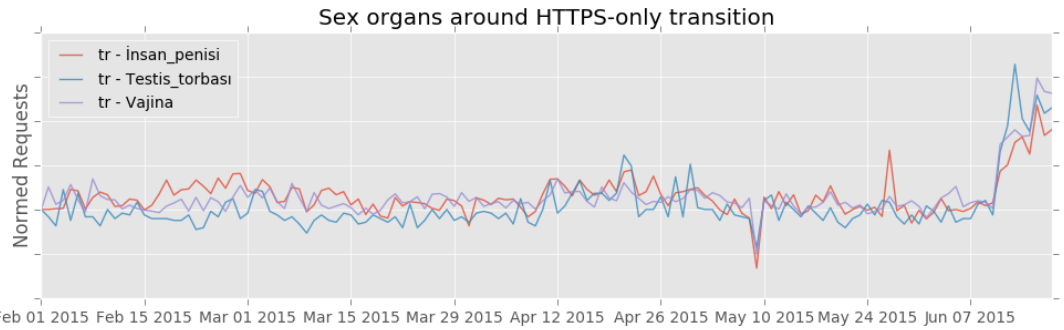
the liability and data retention requirements on hosting and access providers, and add to the list of content considered criminal.[134] [135]

While there are now nine legal categories of criminal content (e.g. content relating to child pornography, obscenity, or gambling), censorship is not limited to these categories. Sites relating to pornography, intellectual property infringement, ethnic minorities, LGBT issues, political movements and news outlets have all been censored.[136] Content unrelated to any of these categories often sees censorship due to the common practice of blocking entire sites for single pieces of infringing content. YouTube, Twitter, Blogger and Wordpress have all been the subject of such blocks,[137] and according to a Turkish watchdog organization, as of June 2016, more than 110,000 unique domains are entirely blocked in Turkey.[138]

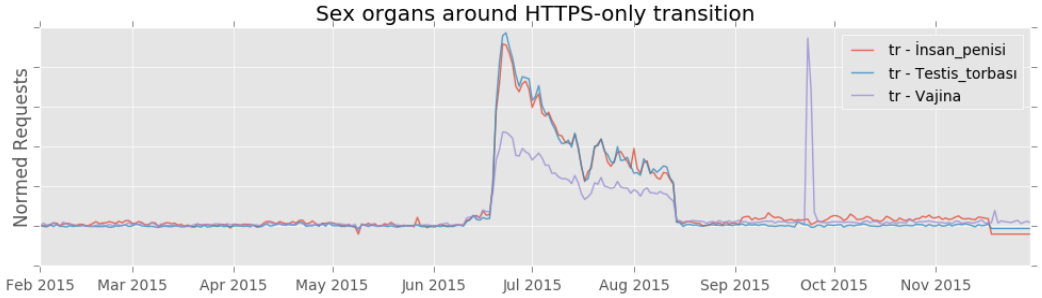
While full domain blocking is widespread, not all censorship occurs at the domain level. Turkey also has the capability to filter individual pages, as has been the case with Turkish Wikipedia. Though it is unclear exactly who ordered the censorship and its full extent,[139] there have been media reports of at least five censored articles: "İnsan penisi" ("Human penis"), "Kadın üreme organları" ("Vulva"), "Testis torbası" ("Scrotum"), "Vajina" ("Vagina"), and "Haziran 2015 Türkiye genel seçimleri için yapılan anketler" ("Opinion polling for the Turkish general election, June 2015").[140] For a number of weeks in the summer of 2015, Turkish Wikipedia included a banner on the main page warning users of this censorship.[141]

Our article-level analysis included data for three of these censored articles:

[134] "Turkey: Freedom on the Net 2014," *Freedom House*, 2014, https://freedomhouse.org/report/freedom-net/2014/turkey.
[135] "Turkey: Freedom on the Net 2015," *Freedom House*, 2015, https://freedomhouse.org/report/freedom-net/2015/turkey.
[136] Ibid.
[137] Mustafa Akgül and Melih Kırlıdoğ, "Internet censorship in Turkey," *Internet Policy Review*, Jun 3, 2015, http://policyreview.info/articles/analysis/internet-censorship-turkey.
[138] Engelli Web, https://engelliweb.com/.
[139] Elif Akgül, "The Anatomy of Vagina Censorship in Turkey," *Bianet,* Nov 18, 2014, http://bianet.org/english/freedom-of-expression/160045-the-anatomy-of-vagina-censorship-in-turkey.
[140] "Wikipedia releases warning on Turkey's censorship, monitoring," *Hurriyet Daily News*, http://www.hurriyetdailynews.com/wikipedia-releases-warning-on-turkeys-censorship-monitoring.aspx?PageID=238&NID=84255&NewsCatID=339.
[141] Internet Archive snapshots of Turkish Wikipedia, *Internet Archive*, Jun 26, 2015 through Aug 10, 2015, http://web.archive.org/web/20150626013525/https://tr.wikipedia.org/wiki/Ana_Sayfa.



At the far right of the above graph, a sharp uptick in the number of requests for each of these three articles is plainly visible. This uptick occurs on June 12, 2015 – the same day as the transition to HTTPS-only delivery. And though we did not limit this analysis to only those requests originating in Turkey, approximately nine out of every ten requests for Turkish Wikipedia come from the country.[142] For these reasons, we believe this is likely an instance of the HTTPS transition enabling more access to these articles. If we look past June 12 though, the picture becomes more complex:



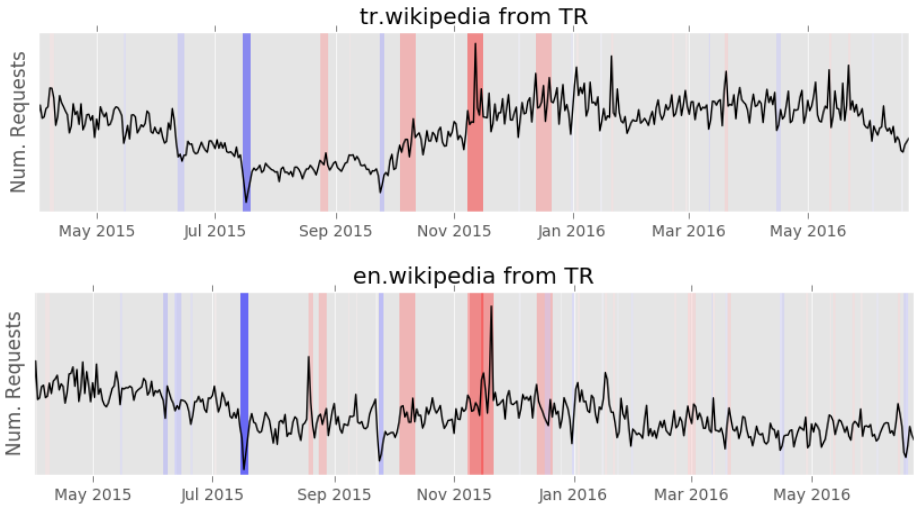
The request spike of June 12 is but a small blip before a much larger increase in traffic on June 19. As can be seen above, this traffic volume slowly decreases over a number of weeks before falling back down around August 14, 2015. It is difficult to attribute this much larger increase and subsequent decrease to any single cause, though we believe automated activity is the most likely culprit. The fact that during this period of increased activity, each article followed a very similar pattern is one piece of evidence pointing to this conclusion. It is interesting to note that the average number of daily requests after this event is higher than the pre-HTTPS average – consistent with the hypothesis that HTTPS enabled more access, regardless of the cause of the intervening anomaly.

Client-side analysis of all Wikipedia projects from a network location in Turkey revealed nothing indicative of domain-level censorship, and round-trip times were within normal ranges:

Median RTT	Mean RTT	Max RTT
128 ms	141.8 ms	948 ms to bxr.wikipedia.org

[142] "Wikimedia Traffic Analysis Report - Page Views Per Wikipedia Language - Breakdown," May 2016, https://stats.wikimedia.org/wikimedia/squids/SquidReportPageViewsPerLanguageBreakdown.htm#Turkish.

More than three-quarters of traffic to Wikipedia from Turkey is bound for Turkish Wikipedia, while much of the rest is directed at English Wikipedia.[143] Neither of these projects showed significant decreases in traffic apart from a short period of time around Ramadan (mid July, 2015):



While all Wikipedia languages projects appeared available in our June 2016 client-side tests from Turkey, our analysis supports the media reports of a number of articles having been blocked in the past for at least a portion of Turkish citizens.

Uzbekistan

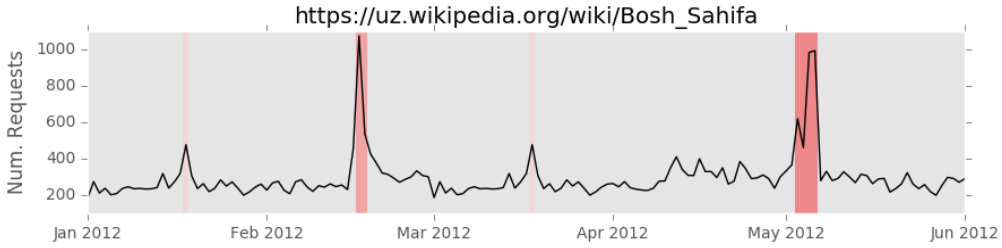
Though it may receive less media attention than other countries with high levels of censorship, Uzbekistan has one of the most intensely controlled online and media environments in the world. Internet censorship has been present in Uzbekistan since about 2002, and has been steadily increasing. Uzbek law prohibits Internet operators from disseminating information that calls for violent overthrow of the government, instigates other forms of violence, is pornographic, relates to religious extremism, or "degrades and defames human dignity."[144] The newly formed government organization that oversees this censorship, the Ministry for the Development of Information Technologies and Communications, is also charged with preventing the "negative influence on the public consciousness of citizens, in particular of young people."[145]

The actual implementation of these laws has created a censorship regime as broad as the legal language suggests. The fairly well-defined categories of pornography and terrorism are blocked, but a host of other topics are censored as well, including: reports of government corruption, human rights organizations (including Amnesty International, Freedom House, and Human Rights Watch),

[143] "Wikimedia Traffic Analysis Report - Wikipedia Page Views Per Country - Breakdown," May 2016, https://stats.wikimedia.org/wikimedia/squids/SquidReportPageViewsPerCountryBreakdown.htm#Turkey.
[144] "Uzbekistan," *OpenNet Initiative*, Dec 21, 2010, https://opennet.net/research/profiles/uzbekistan.
[145] "Uzbekistan: Freedom on the Net 2015," *Freedom House*, Oct 2015, https://freedomhouse.org/report/freedom-net/2015/uzbekistan.

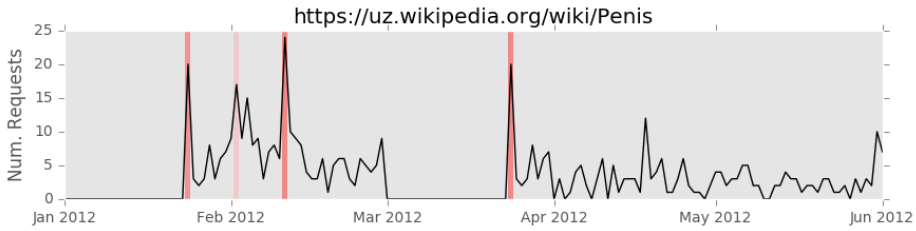
organized crime, political opposition, health education, religious organizations, local NGOs, and independent local and regional news media (including Radio Free Europe/Radio Liberty, Deutsche Welle, and the Uzbek services of the BBC and Voice of America).[146] [147]

Typically, governments target specific URLs for censorship, but occasionally entire domains are blocked.[148] One such occasion was the blocking of all of Uzbek Wikipedia in early 2012.[149] This was an interesting case because Uzbek Wikipedia is less popular in Uzbekistan than either Russian or English Wikipedia, in terms of both articles and the number of requests, and yet neither of those Wikipedias were blocked.[150] An official reason was never given for the ban, but some have speculated that it was due to the addition of a number of articles related to sex that took place shortly before the block.[151] It is possible that only Uzbek Wikipedia was blocked because Uzbek is the only official state language of Uzbekistan.[152] Surprisingly, this block is not visible in the number of requests to Uzbek Wikipedia's main page and was therefore not picked up by anomaly detection:



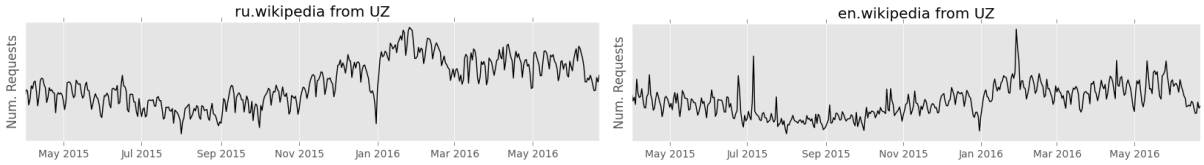
A suspicious drop in requests occurs for articles that were hypothesized to be related to the block, but the date of the reported block and the dates of the anomalies do not agree:

[146] Ibid.
[147] "Uzbekistan," *OpenNet Initiative*, Dec 21, 2010, https://opennet.net/research/profiles/uzbekistan.
[148] Ibid.
[149] "Wikipedia Articles In Uzbek Blocked," *Radio Free Europe/Radio Liberty*, Feb 16, 2012, http://www.rferl.org/content/uzbek_wikipedia_blocked/24486460.html.
[150] "Wikimedia Traffic Analysis Report - Wikipedia Page Views Per Country - Breakdown," May 2016, https://stats.wikimedia.org/wikimedia/squids/SquidReportPageViewsPerCountryBreakdown.htm#Uzbekistan.
[151] Sarah Kendzior, "Why Did Uzbekistan Ban Wikipedia?," *Registan*, Feb 21, 2012, http://registan.net/2012/02/21/why-did-uzbekistan-ban-wikipedia/.
[152] "Uzbekistan," *CIA World Factbook*, Aug 15, 2016, https://www.cia.gov/library/publications/the-world-factbook/geos/uz.html.

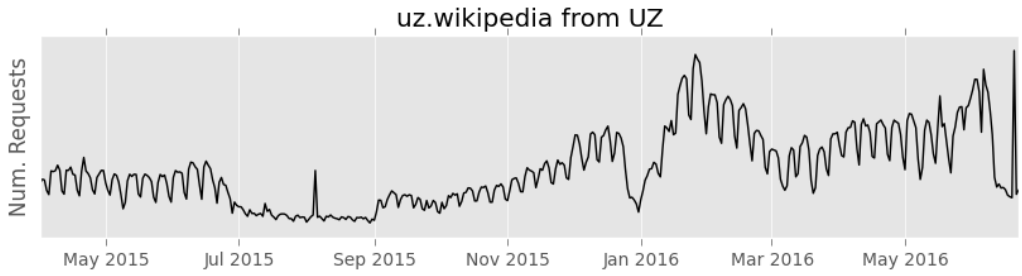


Client side tests from Uzbekistan were significantly different from most of our other client side tests. The biggest difference was that we could not consistently receive successful responses for any Wikipedia project. In fact, requests to all 292 Wikipedia domains returned an error at least once. After further testing, it appeared that this phenomena was not limited to Wikipedia domains—all tested domains intermittently returned errors. This points to an issue in our deployment rather than especially heavy handed censorship. In an effort to continue testing despite likely deployment issues, we repeatedly tested all Wikipedia projects until we received a successful response. For almost all projects, this required between one and three additional tests. After repeated attempts, we were able to successfully request every Wikipedia project except one: Uzbek Wikipedia (http://uz.wikipedia.org/wiki/). Despite a dozen attempts to connect to Uzbek Wikipedia from Uzbekistan over the course of two weeks starting June 25, 2016, we could not receive a single successful response. For this reason, we suspect that Uzbekistan has again started at least partially blocking Uzbek Wikipedia.

Further evidence for this hypothesis can be seen in Wikimedia's server side data. The most popular Wikipedias in Uzbekistan are Russian and English, accounting for about 90% of the traffic. Uzbek Wikipedia makes up only 8.9% of the traffic from Uzbekistan.[153] Starting on June 11, 2016, Russian, English, and Uzbek Wikipedias saw significant drops in the number of requests received. From June 11 through June 17, the average number of daily requests was 19% lower than the previous week for English Wikipedia, 15% lower for Russian Wikipedia, and 62% lower for Uzbek Wikipedia. Unlike previous analyses, this decrease cannot be accounted for by any public holiday that we could identify. These trends can be seen in the following graphs:



[153] "Wikimedia Traffic Analysis Report - Wikipedia Page Views Per Country - Breakdown," May 2016, https://stats.wikimedia.org/wikimedia/squids/SquidReportPageViewsPerCountryBreakdown.htm#Uzbekistan.



Of particular note is the spike that occurred on the far right end of the graph for Uzbek Wikipedia. On June 20, 2016, traffic to Uzbek Wikipedia appeared to return to its previous levels for one day before falling back down to the depressed levels. This spike is not present in requests to Russian or English Wikipedia. While daily spikes of this kind are common in web request data, they are typically spiking above base levels before returning to normal rather than rising quickly to normal levels from depressed levels. This leads us to believe this is unlikely to be a natural traffic event, but rather some external process interfering with requests to Uzbek Wikipedia. As noted above, this would not be unprecedented, though potential motivations for this recent censorship event are unknown. The fact that Uzbek Wikipedia would be blocked but Russian and English Wikipedia would be left untouched could again be explained by the fact that Uzbek is the sole official language of Uzbekistan.[154]

For Wikipedia projects that we could successfully connect to, network round-trip times were quite long:

Median RTT	Mean RTT	Max RTT
8915 ms	8909.7 ms	15208 ms to nrm.wikipedia.org

These long round trip times were not limited to Wikipedia projects and occur across all tests to all URLs, so we believe they were due to our deployment rather than anything related to throttling. The very long wait for nrm.wikipedia.org returned to near average on subsequent tests.

Given the unusual pattern of server-side data and our repeated inability to access Uzbek Wikipedia, we believe it is likely there was some kind of blocking of Uzbek Wikipedia occurring in Uzbekistan as of June 2016.

[154] "Uzbekistan," *CIA World Factbook*, Aug 15, 2016, https://www.cia.gov/library/publications/the-world-factbook/geos/uz.html.

Vietnam

Online activity in Vietnam is tightly restricted through content filtering, fines, website licensing, targeted cyber attacks, and arrests and detentions. The vast majority of content censored in Vietnam is content that could conceivably challenge the power of the ruling political class. In September 2010, OpenNet Initiative researchers found that both of the government-owned ISPs, Viettel and FPT Telecom, were blocking opposition and political reform websites, Vietnamese-language news sites, sites related to the Degar ethnic minority, Facebook, and sites related to circumvention tools.[155] The Decree on Management, Provision, and Use of Internet Services and Information Content Online, adopted in 2013, prohibits the use of the Internet to "oppose the Socialist Republic of Vietnam; threaten the national security, social order, and safety; sabotage the 'national fraternity'; arouse animosity among races and religions; or contradict national traditions."[156] Circular 9, issued in 2014, requires a government license for companies founding new social media sites. The government also employs surveillance, requiring owners of cybercafes to track users' Internet activity.[157]

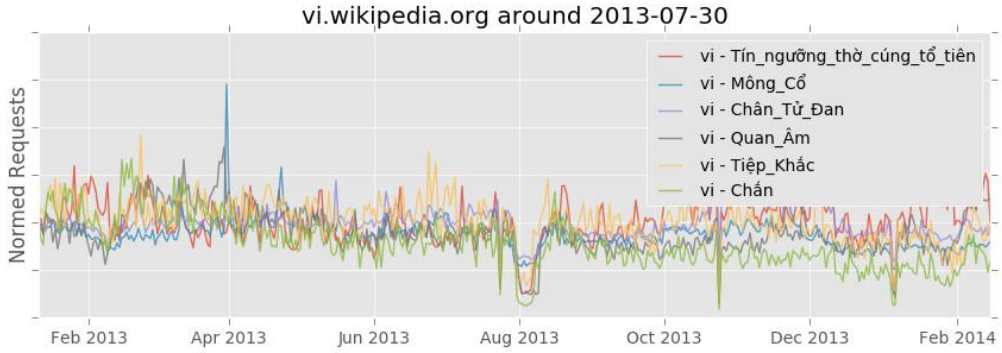
In 2014 and 2015, the government imprisoned 29 bloggers, writers, and activists. Over the past few years, the government has instituted new legislation that strengthens its controls over online content and activity. In 2013, the government issued Decree 72 which forced social media companies to censor content and followed that with Decree 174 in 2014 which authorized punishments for online speech. The Vietnamese government also restricts freedom on the Internet with Article 258 which is a law that bans the "abuse of democratic rights to infringe upon the interests of the State, the legitimate rights and interests of organizations and citizens."[158] In 2014, the government used this law to prosecute over a dozen rights advocates and bloggers and to block two prominent blogs critical of the government. During Obama's official visit in May 2016, the government blocked Facebook to prevent "political dissidents" from voicing their grievances on social media, following a pattern established after it blocked Facebook during environmental protests earlier that month.[159]

We believe that anomalies detected by article-level analysis of Vietnamese Wikipedia are likely associated with requests from Vietnam because almost 90% of traffic to Vietnamese Wikipedia comes from Vietnam.[160] We detected a number of significant anomalies in articles in Vietnamese Wikipedia. For instance, about 130 articles saw significant but temporary drops in traffic around the end of July 2013. These articles do not appear to be thematically related. The following graph

---

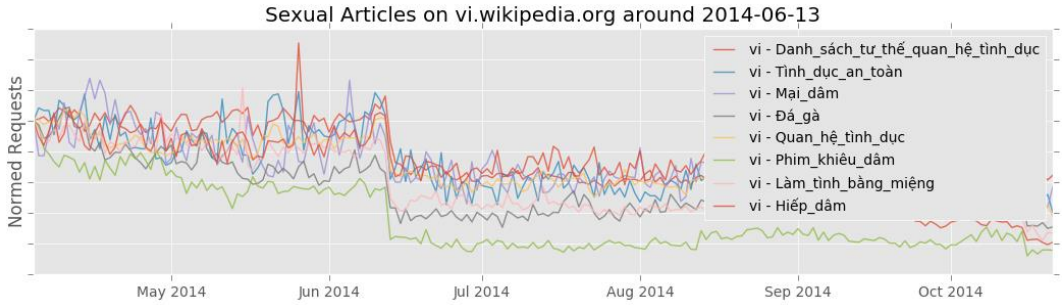
[155] "Vietnam," *OpenNet Initiative*, Aug 7, 2012, https://opennet.net/research/profiles/vietnam.
[156] Sayuri Umedia, "Vietnam: Controversial Internet Decree in Effect," *Global Legal Monitor*, Sep 6, 2013, http://www.loc.gov/law/foreign-news/article/vietnam-controversial-internet-decree-in-effect/.
[157] "Vietnam: Freedom on the Net 2015," *Freedom House*, Oct 2015, https://freedomhouse.org/report/freedom-net/2015/vietnam.
[158] Ibid.
[159] Jessica Conditt, "Activists Say Vietnam shut down Facebook during Obama's visit," *Engadget*, May 27 2016, https://www.engadget.com/2016/05/27/vietnam-obama-facebook-blocked-activists-politics/.
[160] "Wikimedia Traffic Analysis Report - Page Views Per Wikipedia Language - Breakdown," May 2016, https://stats.wikimedia.org/wikimedia/squids/SquidReportPageViewsPerLanguageBreakdown.htm#Vietnamese.

depicts "Veneration of the dead," "Mongolia," "Donnie Yen," "Guanyin," "Czechoslovakia," and "Chắn" (a card game):



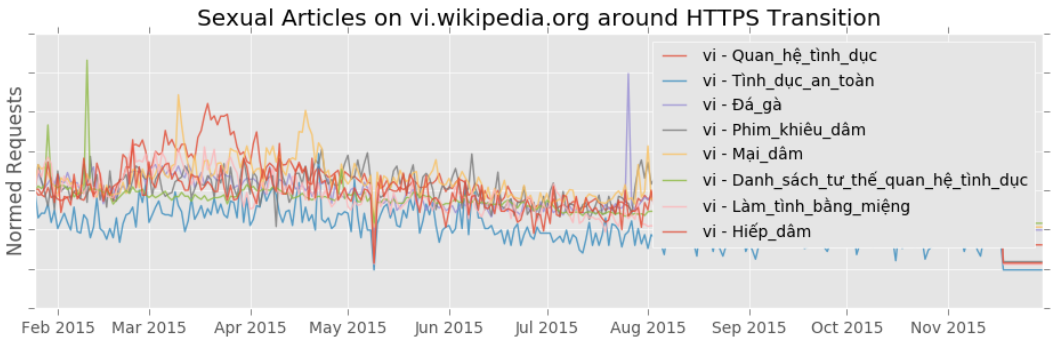
We were unable to find a suitable explanation for this drop in traffic, either internal or external to Wikipedia.

We did detect an anomaly for a number of articles that appear thematically related beginning around June 13, 2014. All of the articles to which traffic decreased substantially around this time are related to sex with the exception of "Đá gà," which translates to "Cockfight." We were unable to associate this decrease with any change internal to Wikipedia (e.g., the move or deletion of the article).



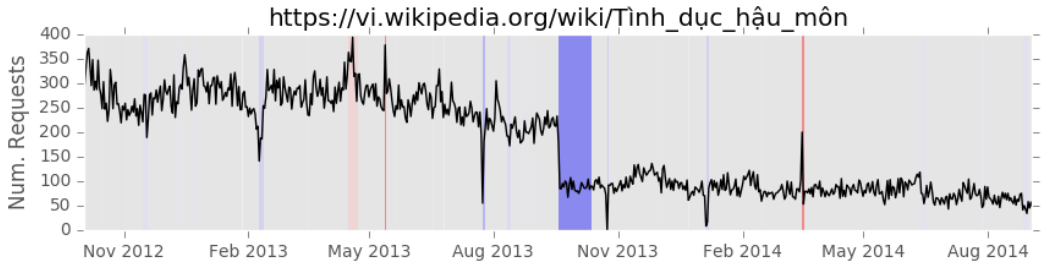
Article	Translation
Làm tình bằng miệng	Oral sex
Đá gà	Cockfight
Quan hệ tình dục	Sexual intercourse
Phim khiêu dâm	Pornographic film
Danh sách tư thế quan hệ tình dục	Sex position
Mại dâm	Prostitution
Tình dục an toàn	Safe sex
Hiếp dâm	Rape

Note that unlike in the Iranian case, traffic to these articles does not appear to increase substantially after June 12, 2015:



It is unclear why we do not see a concomitant increase in traffic as in the Iranian case, but there are at least two plausible explanations that are consistent with the observed data. First, the decrease in traffic could be unrelated to censorship and instead be due to some other change (e.g. cessation of bot activities) that would not be affected by HTTPS. This would be consistent with the fact that while pornography is illegal in Vietnam,[161] there are no known instances of censorship of Internet pornography. Second, it could be the case that this was a censorship event but user behavior toward the censored pages changed in such a way that traffic did not increase once the pages were available again (e.g. linking patterns changed in the intervening period to route away from the censored articles).

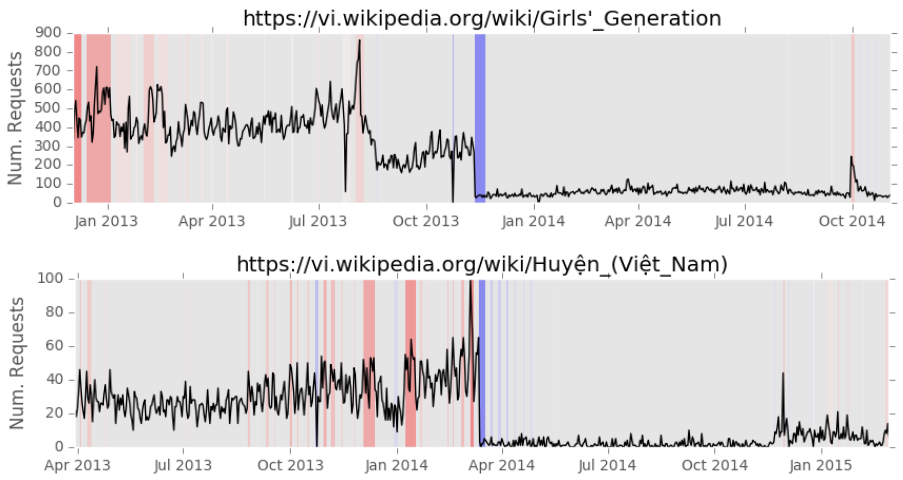
Perhaps relatedly, this larger event was predated by a significant decrease in the amount of traffic to the article "Tình dục hậu môn" ("Anal sex") on September 18, 2013:



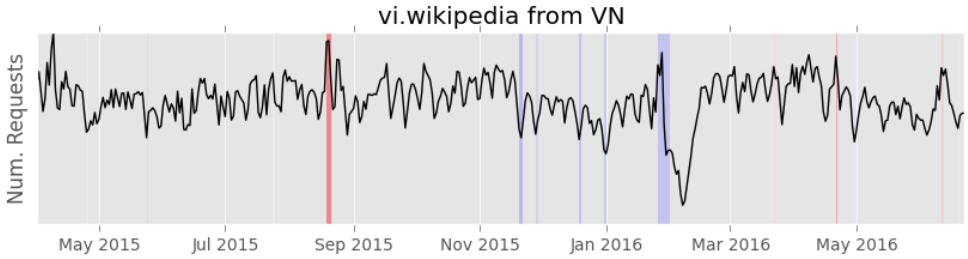
Again, this change does not appear to be associated with any change internal to Wikipedia. Unlike the previous event, it is not correlated in time with any other significant anomalies of the same nature.

Other articles that do not appear explicitly sexual in nature also experienced significant decreases in traffic that cannot be accounted for by actions on Wikipedia alone. These include the articles for "Girls' Generation," a South Korean music group, and "Huyện (Việt Nam)," an administrative district in Vietnam:

[161] "Vietnam," *Australian Department of Foreign Affairs and Trade*, Aug 18, 2016, http://smartraveller.gov.au/countries/asia/south-east/pages/vietnam.aspx.



Project-level traffic from Vietnam to Vietnamese Wikipedia appeared normal from May 2015 to July 2016 with the exception of the beginning of February, which is the multi-day holiday of Vietnamese New Year (Tết).



Our client-side testing from Vietnam showed no problems accessing any Wikipedia project subdomains, and page load times were within normal ranges:

Median RTT	Mean RTT	Max RTT
367 ms	382.2 ms	980 ms to bxr.wikipedia.org

While not conclusive on its own, we believe we have surfaced tentative evidence that at least some portion of Vietnam's Internet users may have been blocked from accessing sexually explicit articles in the past. The HTTPS transition now makes this type of censorship unlikely, and our client and server data analysis of June 2016 showed no evidence that Vietnam was blocking the entirety of any Wikipedia project.

Additional Findings

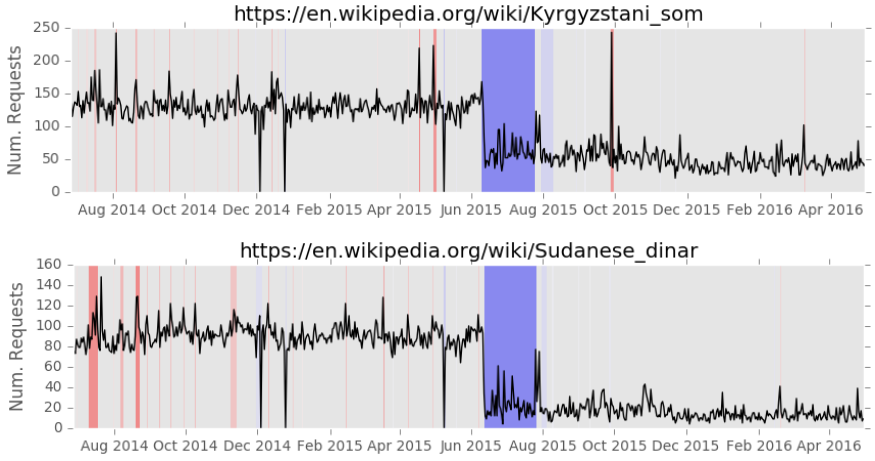
Each of our analysis methods had additional results that did not pertain directly to the countries enumerated above. We have provided these results here, organized by our method of analysis.

Article-level Analysis

Overall, our article analysis pipeline detected 92.4 million anomalies across the 1.7 million articles. We started by looking at the anomalies that indicated significant drops in traffic, and the first things we noticed were nine distinct periods of time in which traffic dropped precipitously for a large number of articles across most, if not all, Wikipedia projects. These events were fairly short, most lasting only a day. We believe these were likely periods of faulty data collection. The spring of 2013 also saw many articles across many projects lose significant traffic; these events were harder to explain as they did not all begin at the same time and often lasted for weeks. Together, the nine data collection errors and the spring of 2013 accounted for a large number of the most significant downward anomalies we witnessed. Due to their widespread nature, we considered them unlikely to be related to censorship, and therefore excluded them from the rest of our analysis. These periods are outlined in more detail in Appendix D.

After removing these dates, we were left with 84.1 million anomalies. 71 million were anomalous increases in traffic (which our pipeline also detected but we only briefly reviewed) and the remaining 13.1 million anomalies were significant decreases in traffic.

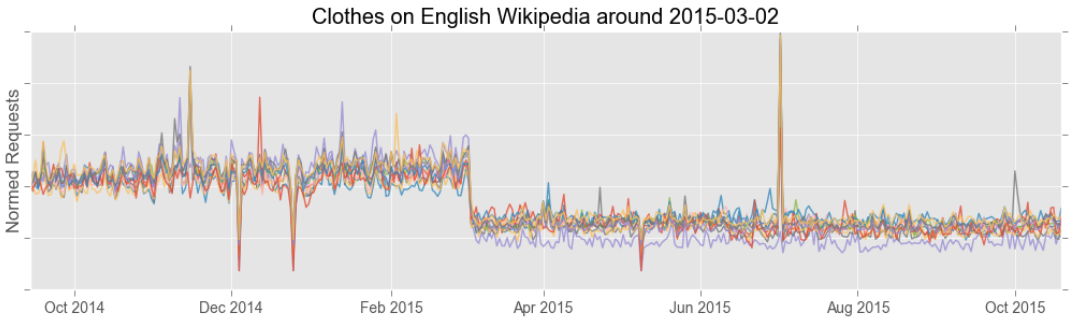
One more date of note was the date Wikipedia changed over to HTTPS for all traffic. While this might not have taken place simultaneously across the globe, it looks as though June 12, 2015 was most common date. Many, though not all, articles saw significant drops in traffic around this date. Eighteen of the articles that saw the sharpest decreases were articles for various currencies on English Wikipedia:



While we were unsure of the cause, a plausible explanation might be that the infrastructural change that the transition required affected the collection of the request metrics to some degree. The data that we used for our analysis did not have requests by bots or spiders filtered out, so it is also possible that these automated processes were using HTTP to request articles and could not handle the redirect to HTTPS. It could also be the case that some network operators were performing full HTTPS protocol blocking. Whatever the cause, per-project request metrics also saw a number of drops around the HTTPS transition. The graph of Fiji below in the Server-side Data section is a good example.

Because our article-level analysis was broken out by language rather than country, we have a number of results that are limited to a single language, but could potentially relate to one or more countries. For example, nine countries each contribute more than 1% of the requests to English Wikipedia, including countries known to have blocked articles in the past (e.g., Iran).[162] This makes it difficult to attribute anomalies to any individual country for languages like English that are spoken in many countries. Nevertheless, we felt it important to include these anomalies for both completeness and to aid any research that may follow this report. Though we located and include a number of these anomalies, we did not spend as much effort investigating anomalies for these languages.

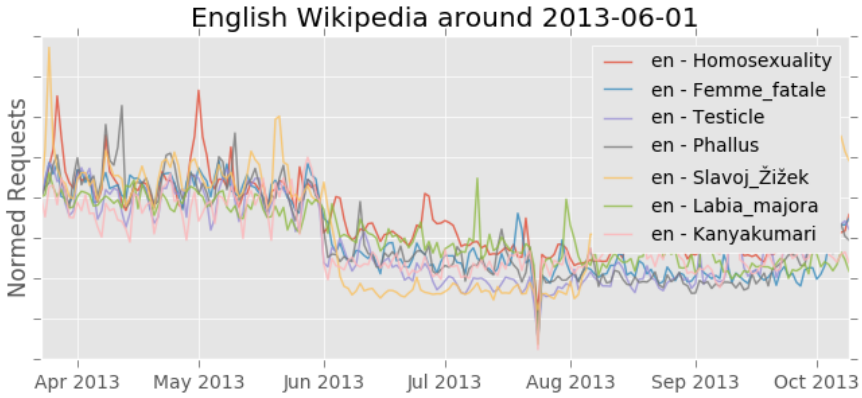
One event that stood out on English Wikipedia was a cluster of anomalies beginning around March 2, 2015. Twelve articles, all related to clothing, saw the most significant drops in traffic around that date. Those twelve articles are: "Blanket sleeper," "Coin purse," "Débutante dress," "Denim skirt," "Goggle jacket," "Gymslip," "Jodhpurs," "Nightshirt," "Nightwear," "Opera coat," "Swim diaper," and "Undershirt." A graph of their overlapping time series shows that the patterns are remarkably similar:



Given the fairly innocuous or archaic natures of many of these articles, we think this is likely a good example of some automated process ceasing operations that previously contributed a large share of requests.

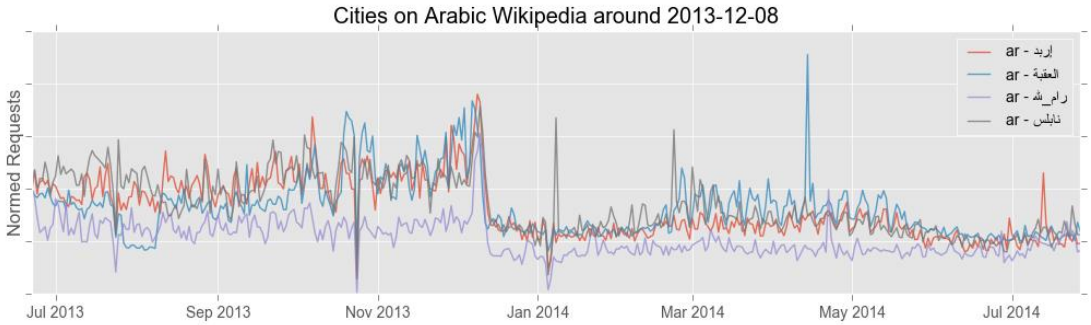
[162] "Wikimedia Traffic Analysis Report - Page Views Per Wikipedia Language - Breakdown," May 2016, https://stats.wikimedia.org/wikimedia/squids/SquidReportPageViewsPerLanguageBreakdown.htm#English.

Less innocuous articles saw correlated decreases in traffic around the beginning of June 2013. This event covered at least the following articles: "Femme fatale," "Hermaphrodite," "Homosexuality," "Kanyakumari," "Labia majora," "Mario," "Pantyhose," "Phallus," "Slavoj Žižek," "Testicle," "Undergarment." The following graph depicts this trend:



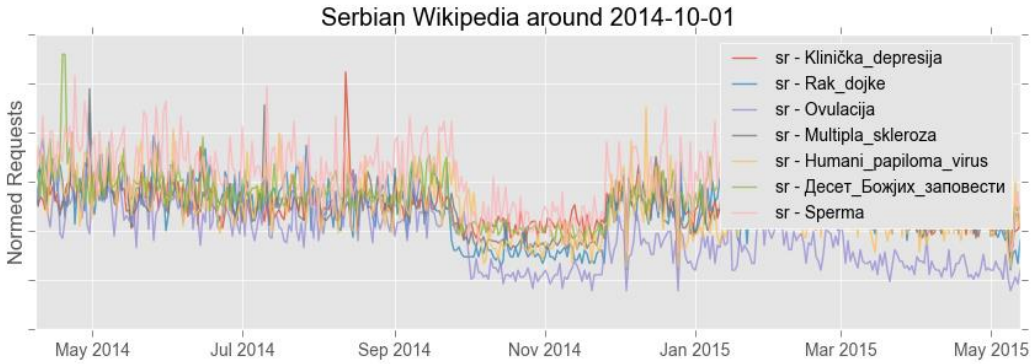
Nothing was discovered in the histories of these articles that could explain their sudden drop in traffic.

Arabic Wikipedia saw a number of significant anomalies, but after investigation, a large share of these anomalies were likely caused by Wikipedia users changing article titles or introducing redirects. We did identify one cluster of thematically related anomalies. Beginning near December 8, 2013, traffic to articles for four Middle Eastern cities saw a quick increase and then sharp and sustained drop off. The following graph illustrates the phenomenon:



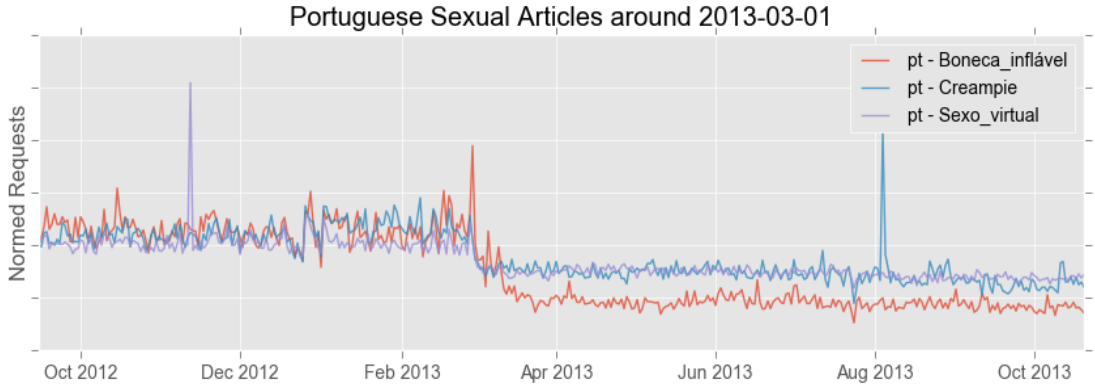
These cities translate to, in the order shown in the graph's legend, "Irbid," "Aqaba," "Ramallah," and "Nablus," Due to relatively innocuous nature of the content and the synchronized increase in traffic across the four articles that we could not attribute to outside factors, we suspect this might be related to bot activity rather than censorship.

We did not cover the country of Serbia, but Serbian Wikipedia saw a temporary drop in requests to a number of articles that are mostly related to human health:

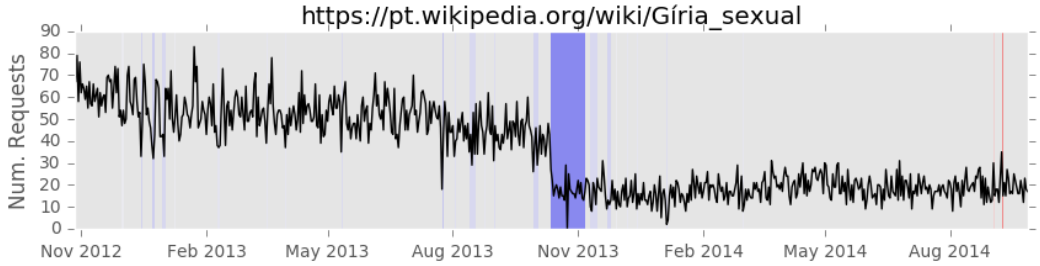


The English equivalents of these articles, in the order they are given in the graph's legend, are "Clinical depression," "Breast cancer," "Ovulation," "Multiple sclerosis," "Human papillomavirus," "Ten Commandments," and "Sperm." Nothing in the histories of these articles suggests a cause for this anomaly.

Sex-related articles on Portuguese Wikipedia saw a number of significant downward anomalies, but three articles related to sex all saw anomalies at around the same time that cannot be explained by changes in Wikipedia alone: "Boneca inflável" ("Sex doll"), "Creampie" ("Creampie (sexual act)"), and "Sexo virtual" ("Virtual sex"). These anomalies took place around March 1, 2013, and are seen below:



"Gíria sexual" ("Sexual slang") saw a similar anomalous event, though months later on October 12, 2013:

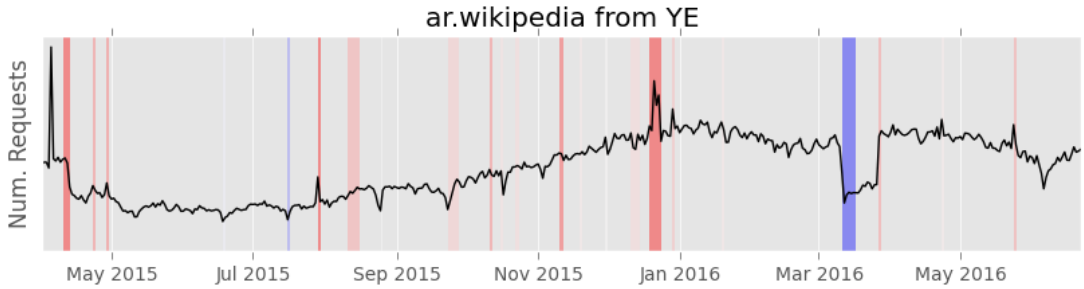


We felt that though suspicious, the small number of anomalies meant there was not enough evidence to suggest censorship.

Project-level Analysis

Anomaly detection on the number of daily requests at the project level turned up a number of interesting decreases in traffic from a number of different countries. We did not analyze all of the detected anomalies, but we did investigate a number of the larger anomalies in an effort to find potential causes. While causal links are hard to establish, there is evidence to suggest that inaccessibility caused by war, governmental decree, and natural disaster are all detectable in Wikipedia's data.

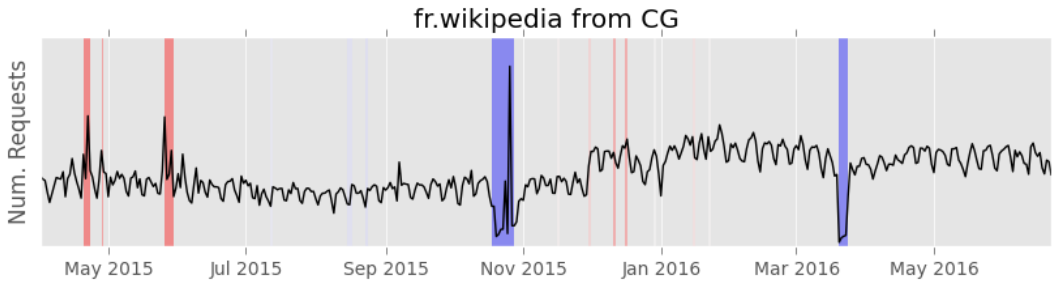
One of the largest anomalies our analysis uncovered was a decrease in traffic from Yemen that lasted for more than two weeks. From March 11 until March 27, 2016, requests for Arabic Wikipedia were down approximately 50% from normal levels. The same holds true for English Wikipedia. The anomaly is easily visible in the traffic graph:



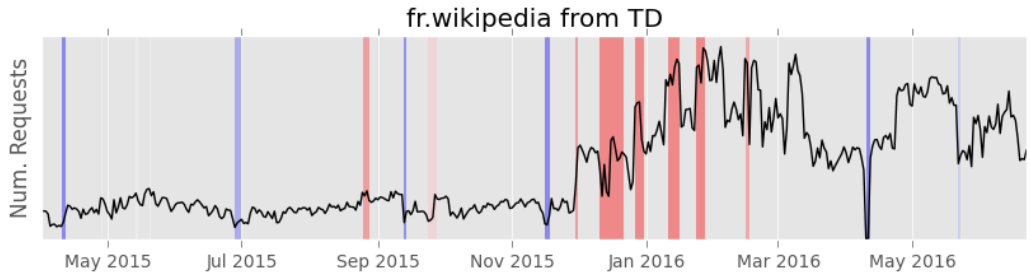
We believe this event might have been related to fighting between the Yemeni government and rebel groups that could have caused infrastructure outages, though we did not find media reports of such outages. This anomaly occurs around the same time fighting intensified in the country as the government forces broke the rebels' siege of Taiz, Yemen's third largest city.[163]

The Republic of the Congo saw two fairly large decreases in traffic to French Wikipedia, the largest recipient of the country's traffic.[164] On October 20, 2015, a large, multi-day anomaly began around the same time as protests against the country's president.[165] Months later, when a presidential election took place, the country ordered a total media blackout.[166] This outage is obvious when looking at a graph of the data:

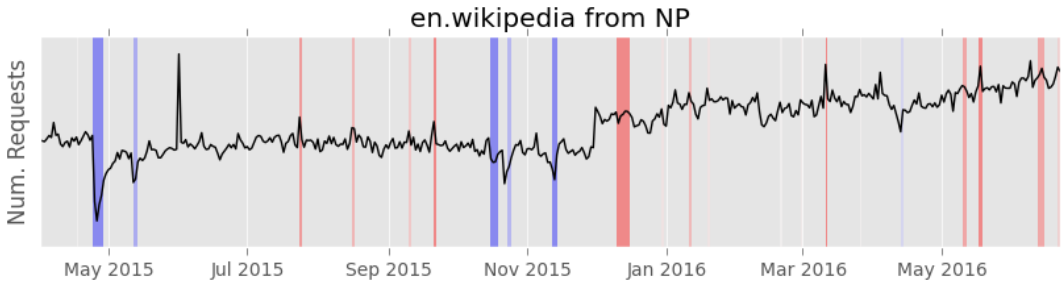
[163] "Yemeni government gains ground in besieged Taiz," *Al Jazeera*, Mar 12, 2016, http://www.aljazeera.com/news/2016/03/yemeni-government-gains-ground-besieged-city-taiz-160311210932533.html.
[164] "Wikimedia Traffic Analysis Report - Wikipedia Page Views Per Country - Breakdown," May 2016, https://stats.wikimedia.org/wikimedia/squids/SquidReportPageViewsPerCountryBreakdown.htm#Congo Rep..
[165] Lily Kuo, "Congo Brazzaville is in an uproar against its would-be president-for-life," *Quartz*, Oct 20, 2015, http://qz.com/528516/congo-is-in-an-uproar-against-its-would-be-president-for-life/.
[166] "Congo in media blackout for presidential elections," *Al Jazeera*, Mar 20, 2016, http://www.aljazeera.com/news/2016/03/congo-media-blackout-presidential-elections-160320044041238.html.



A similar tactic seems to have been used by the government of Chad immediately after their presidential elections on April 10, 2016:[167]



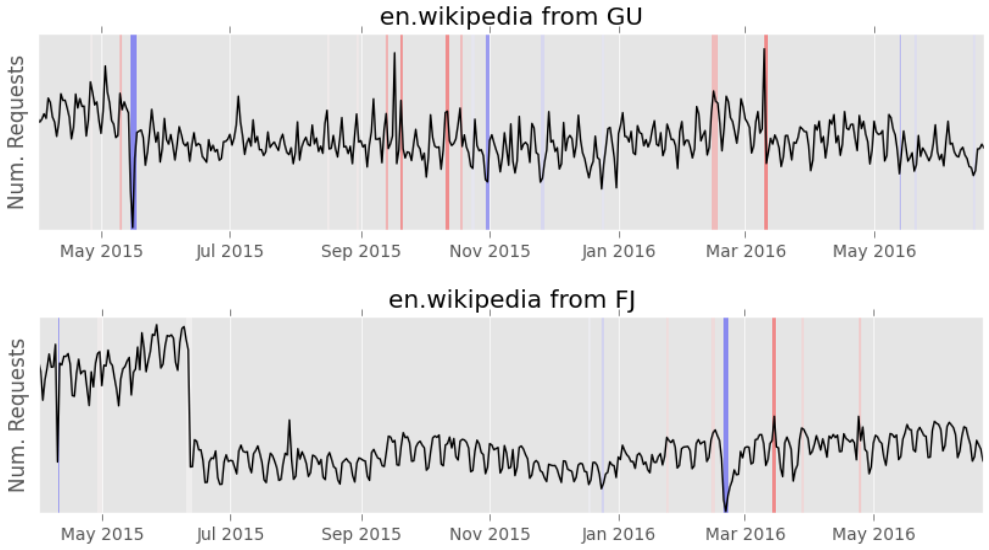
Natural disasters likely accounted for a large number of traffic anomalies detected by our algorithms. Perhaps the largest natural disaster that was seen in the data was the earthquake centered near Katmandu, Nepal on April 24, 2015. The earthquake caused extensive damage in the capital, surely knocking out Internet access along with power to a significant percentage of the population.[168] The outage and the recovery can be seen toward the left side of the following graph:



[167] "Internet cut in Chad after tense polls," *News24*, Apr 11, 2016, http://www.news24.com/Africa/News/internet-cut-in-chad-after-tense-polls-20160411.

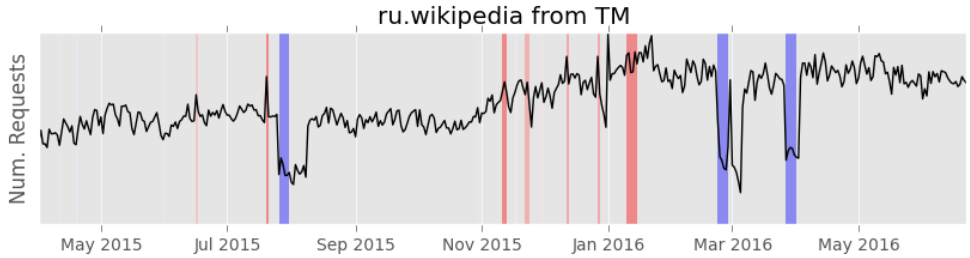
[168] Ellen Barry, "Earthquake Devastates Nepal, Killing More Than 1,900," *New York Times*, Apr 25, 2015, http://www.nytimes.com/2015/04/26/world/asia/nepal-earthquake-katmandu.html?_r=0.

Guam and Fiji both experienced typhoons that were likely the causes of distinct drops in traffic on May 15, 2015 and February 20, 2016, respectively:[169][170]



The large drop in traffic that appears in the left of the Fiji graph around the middle of June 2015 coincides with Wikipedia's switch to HTTPS-only delivery.

Finally, traffic from Turkmenistan saw a number of large anomalies that we found difficult to explain. These anomalous events began on July 26, 2015, February 23, 2016, and March 27, 2016:



Client-side Analysis

In addition to the results already described above, we were able to access all Wikipedia projects in reasonable amounts of time from network locations in the following countries: Algeria, Armenia, Azerbaijan, Bangladesh, Belarus, Brazil, Bulgaria, Cambodia, Canada, Czech Republic, Hong Kong, Hungary, India, Israel, Italy, Jordan, Kuwait, Kyrgyzstan, Lebanon, Malaysia, Moldova, Morocco, Netherlands, Palestine, South Africa, Spain, Turkey, Ukraine, and the United Arab Emirates.

[169] "Residents Take Shelter as Typhoon Pelts Guam With Winds, Rain," *NBC News*, May 15, 2015, http://www.nbcnews.com/news/world/residents-take-shelter-typhoon-pelts-guam-winds-rain-n359911.
[170] Kevin Conlon, Joshua Berlinger and Ralph Ellis, "In Fiji, 17 dead from 'monster' Cyclone Winston; schools shuttered for a week," *CNN*, Feb 22, 2016, http://www.cnn.com/2016/02/21/asia/fiji-tropical-cyclone-winston.

Next Steps and Conclusions

This report is part of a larger project aimed at locating the global boundaries of access to Wikipedia. One of the weaknesses of the format is that reports are inherently locked in time while the Internet censorship landscape continues to change. To complement this report in ways that do not have the same drawback, we are currently undertaking two efforts: continued client-side availability monitoring and continued server-side data monitoring. Client-side monitoring of Wikipedia's projects will continue as long as resources allow. In the short term, it is due to expand as new vantage points are scheduled to come online in the latter half of 2016 that were not available during the writing of this report. Server-side monitoring of the levels of traffic from various countries to Wikipedia's language projects will continue in the medium-term and significant anomalies will be brought to the attention of the Wikimedia Foundation. Manual analysis and investigation of the detected anomalies for the purpose of informing the Wikimedia Foundation of potential censorship will not continue as the process is currently intensive in terms of both time and resources.

We believe that with support and further development, the process of detecting censorship and other outage events from Wikipedia data could be further automated and significantly improved. We have a number of ideas in this vein, some of which could leverage existing Wikipedia research.[171]

While our process does not yet scale to the full size of Wikipedia, we believe that our multimodal methodology—and anomaly detection in particular—has real, demonstrable value to a number of communities. First, we hope that the research in this report has some utility to the Wikimedia Foundation in their efforts to make all knowledge freely available to every person. Second, in the process of doing this research, we have created a dataset of anomalies in request traffic to a select number of individual articles. This dataset is likely useful in answering research questions around Wikipedia itself, but it and others like it could be used to answer questions around singular and significant events in the demand for specific pieces of knowledge. We will publish our generated dataset and open source our anomaly detection pipeline.[172] Third, with Wikipedia's vast size and millions of daily requests, the Wikimedia Foundation has an incredible vantage point to witness events around the Internet beyond even the scope of its own large projects. We have shown that Wikipedia's data can be used to discover and track Internet shutdowns and broader outages around the world. If developed into a publicly accessible resource, this could be a tremendous data source for those interested in Internet accessibility issues.

While some of the raw data might be difficult to publish in a way that still preserves privacy, publishing the anomalies detected in Wikipedia's data has far fewer privacy concerns. As outlined

[171] For example, determining thematic clusters of anomalies could be aided by some of the semantic analysis research that has been done or is actively ongoing (e.g. https://meta.wikimedia.org/wiki/Research:Wikipedia_Navigation_Vectors or https://meta.wikimedia.org/wiki/Research:Investigating_Semantic_Navigation_on_Wikipedia).
[172] The open sourced anomaly detection pipeline software is available at https://github.com/berkmancenter/hekaanom.

above, anomalies at both the article and project level could still have extraordinary value to researchers, advocates, political scientists, sociologists, media and communication scholars, developers of circumvention technologies, policymakers, and others. The generation and publication of this data would also be a fine addition to the Wikimedia Foundation's mission to accumulate and share knowledge.

This information could also help alert the world of those interfering with the Wikimedia Foundation's mission. As of June 2016, it appears China, Thailand, and Uzbekistan are all likely interfering with or completely censoring some part of Wikipedia. The evidence we collected suggests that in each case, the censorship is limited to a single project (Chinese, Yiddish, and Uzbek Wikipedias, respectively). While collectively these projects contain more than one million articles,[173] considering the widespread use of filtering technologies and the vast coverage of Wikipedia, there is currently relatively little censorship of Wikipedia globally. In fact, our research suggests that on balance, there is less censorship happening now than before the transition to HTTPS-only content delivery in June 2015. This initial data suggests the decision to shift to HTTPS has been a good one in terms of ensuring accessibility to knowledge.

And though the current level of censorship may be relatively low, in an ideal world, the Wikimedia Foundation would need not tolerate any censorship. When working toward that ideal world, there are many priorities and values to balance. We hope that our research has provided some useful context and a number of possible options to consider as the Wikimedia Foundation advances its mission in the future.

[173] "List of Wikipedias," *Wikimedia Foundation Labs*, http://wikistats.wmflabs.org/display.php?t=wp.

Appendix A: Wikipedia Projects

The list of the 292 Wikipedia projects assessed by this report, designated by their subdomains, is as follows:

aa, ab, ace, ady, af, ak, als, am, ang, an, arc, ar, arz, ast, as, av, ay, azb, az, bar, bat-smg, ba, bcl, be-tarask, be, bg, bh, bi, bjn, bm, bn, bo, bpy, br, bs, bug, bxr, ca, cbk-zam, cdo, ceb, ce, cho, chr, ch, chy, ckb, co, crh, cr, csb, cs, cu, cv, cy, da, de, diq, dsb, dv, dz, ee, el, eml, en, eo, es, et, eu, ext, fa, ff, fiu-vro, fi, fj, fo, frp, frr, fr, fur, fy, gag, gan, ga, gd, glk, gl, gn, gom, got, gu, gv, hak, ha, haw, he, hif, hi, ho, hr, hsb, ht, hu, hy, hz, ia, id, ie, ig, ii, ik, ilo, io, is, it, iu, ja, jbo, jv, kaa, kab, ka, kbd, kg, ki, kj, kk, kl, km, kn, koi, ko, krc, kr, ksh, ks, ku, kv, kw, ky, lad, la, lbe, lb, lez, lg, lij, li, lmo, ln, lo, lrc, ltg, lt, lv, mai, map-bms, mdf, mg, mhr, mh, min, mi, mk, ml, mn, mo, mrj, mr, ms, mt, mus, mwl, myv, my, mzn, nah, nap, na, nds-nl, nds, ne, new, ng, nl, nn, nov, no, nrm, nso, nv, ny, oc, om, or, os, pag, pam, pap, pa, pcd, pdc, pfl, pih, pi, pl, pms, pnb, pnt, ps, pt, qu, rm, rmy, rn, roa-rup, roa-tara, ro, rue, ru, rw, sah, sa, scn, sco, sc, sd, se, sg, sh, simple, si, sk, sl, sm, sn, so, sq, srn, sr, ss, stq, st, su, sv, sw, szl, ta, tet, te, tg, th, ti, tk, tl, tn, to, tpi, tr, ts, tt, tum, tw, tyv, ty, udm, ug, uk, ur, uz, vec, vep, ve, vi, vls, vo, war, wa, wo, wuu, xal, xh, xmf, yi, yo, za, zea, zh-classical, zh-min-nan, zh, zh-yue, zu

Note that ii.wikipedia.org (the Nuosa language) consistently returned a 404 HTTP status code.

Appendix B: Client Test Details

The following is the list of countries from which we tested each Wikipedia project's subdomain:

Algeria, Armenia, Azerbaijan, Bahrain, Bangladesh, Belarus, Brazil, Bulgaria, Cambodia, Canada, China (2 nodes), Czech Republic, Egypt, Hong Kong, Hungary, India, Indonesia, Israel, Italy, Jordan, Kazakhstan, Kuwait, Kyrgyzstan, Lebanon, Malaysia, Moldova, Morocco, Netherlands, Pakistan, Palestinian Territory, Russian Federation, Saudi Arabia, South Africa, South Korea, Spain, Thailand, Turkey, Ukraine, United Arab Emirates, Uzbekistan, Vietnam

To test from these locations, we partnered with a third-party who has access to a number of servers around the world. Most of these servers, and all the servers utilized in this report, were either co-located servers or virtual private servers rented by our partner from individual companies on the ground in each country. Our partner supplied us with an API that allowed us to issue requests and retrieve data from each of these servers. All steps that a typical client would perform were performed from these client servers: resolve the URL to IP address using a designated DNS server, request HTTP resources from the resolved IP, follow any HTTP redirects, render all returned resources in a WebKit-based browser, etc.

While the API allowed us to specify the DNS servers that should be used by the in-country servers to perform the request, we did not have a list of in-country DNS servers. While in-country open DNS servers are not difficult to locate,[174] we have ethical concerns around causing DNS servers that may be located in private homes to issue recursive requests for potentially censored domains. For that reason, we did not include DNS servers in our requests, and instead DNS resolution fell back to the default DNS servers, Google's Public DNS (8.8.8.8 or 8.8.4.4). This means that any censorship that was only implemented as DNS tampering (as opposed to IP-, TCP-, or HTTP-based censorship) would not have been detected by our client tests. While DNS tampering is widespread,[175] it is not clear how often DNS tampering is the sole method of censorship. Iran, for example, uses DNS tampering as only one of several methods.[176] We intend to locate ISP-supplied in-country DNS servers in the future.

[174] Multiple services exist that provide the IP address and geographic location of DNS servers, e.g. https://censys.io/ or https://www.shodan.io/.
[175] Stéphane Bortzmeyer, "DNS Censorship (DNS Lies) As Seen By RIPE Atlas," *RIPE*, Dec 11, 2015, https://labs.ripe.net/Members/stephane_bortzmeyer/dns-censorship-dns-lies-seen-by-atlas-probes.
[176] Simurgh Aryan et al., "Internet Censorship in Iran: A First Look," August 2013, https://jhalderm.com/pub/papers/iran-foci13.pdf.

Appendix C: Likely Censored Persian Articles

The following is the list of articles on Persian Wikipedia that we identified as increasing in traffic substantially after Wikipedia's transition to solely HTTPS:

Article	English Equivalent
آمیزش_جنسی	Sex
فرج‌لیسی	Cunnilingus
طبق‌زنی	Tribadism
مهبل	Vagina
سکس_شاپ	Sex_shop
تن‌فروشی	Prostitution
صنعت_سکس	Sex_industry
برانگیختگی_جنسی	Sexual_arousal
سید_محمد_خاتمی	Mohammad_Khatami
تحریک_دهانی_نوک_پستان	Erotic_lactation
یادگارخواهی	Sexual_fetishism
سپاه_پاسداران_انقلاب_اسلامی	Army_of_the_Guardians_of_the_Islamic_Revolution
تجاوز_جنسی	Rape
ختنه‌گاه	Circumcision_scar
والایش	Sublimation_(psychology)
عبدالله_نوری	Abdollah_Nouri
سازمان_مجاهدین_انقلاب_اسلامی	Mojahedin_of_the_Islamic_Revolution_Organization
مجمع_روحانیون_مبارز	Association_of_Combatant_Clerics
بی‌دی‌اس‌ام	BDSM
جبهه_مشارکت_ایران_اسلامی	Islamic_Iran_Participation_Front
اکبر_گنجی	Akbar_Ganji
سوماتا	Sumata

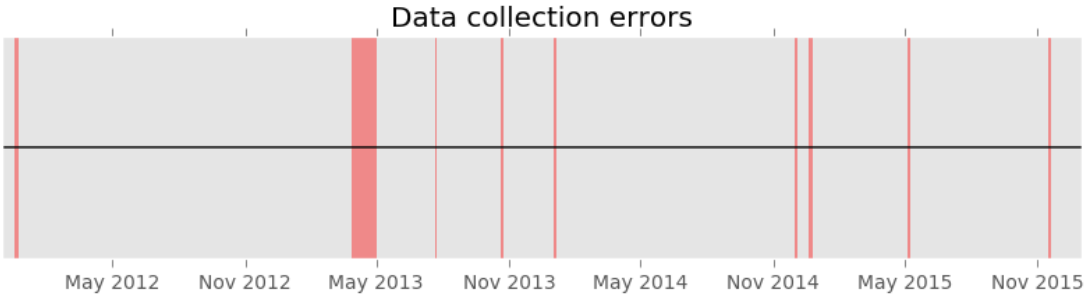
We reviewed each of these articles for edits in their history that could explain an increase in traffic (such as a more popular article getting retitled) and found nothing in each case. We did not understand the article history of "صنعت_سکس" ("Sex_industry"), for which we had historical request data going back years, but whose history showed article creation in February of 2016.[177] The Wikipedia log did not show any prior deletions of the article.[178]

[177] "صنعت سکس," *Persian Wikipedia*, https://fa.wikipedia.org/w/index.php?title=صنعت_سکس&action=history.
[178] "تمام سیاهه‌های عمومی," *Persian Wikipedia*, https://fa.wikipedia.org/w/index.php?title=ویژه%3Aسیاهه‌ها&type=&user=&page=صنعت_سکس.

Appendix D: Dates with Widespread Anomalies

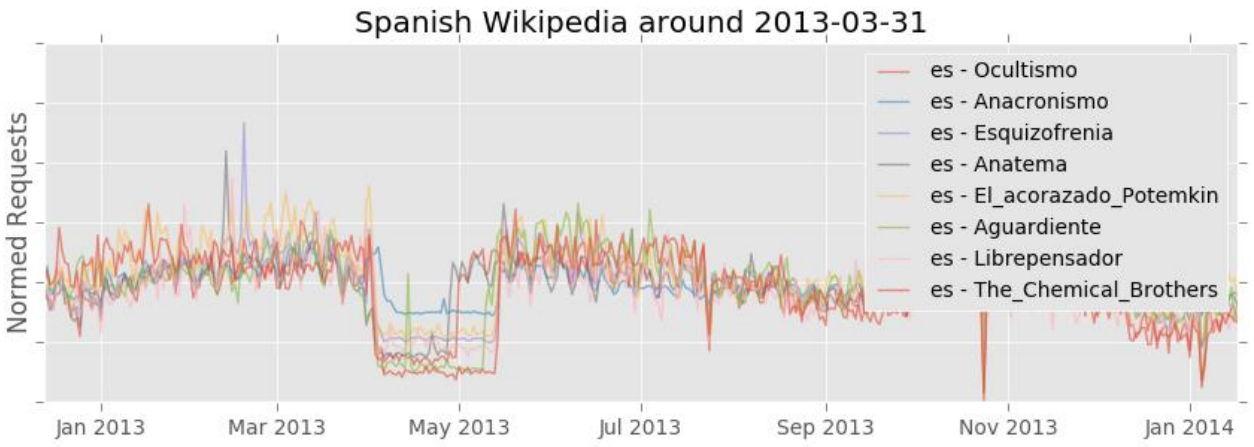
The following time spans were considered likely data collection errors, and were therefore excluded from analysis. In each of these time ranges, traffic to many articles across many projects fell to zero or near zero before returning to normal levels:

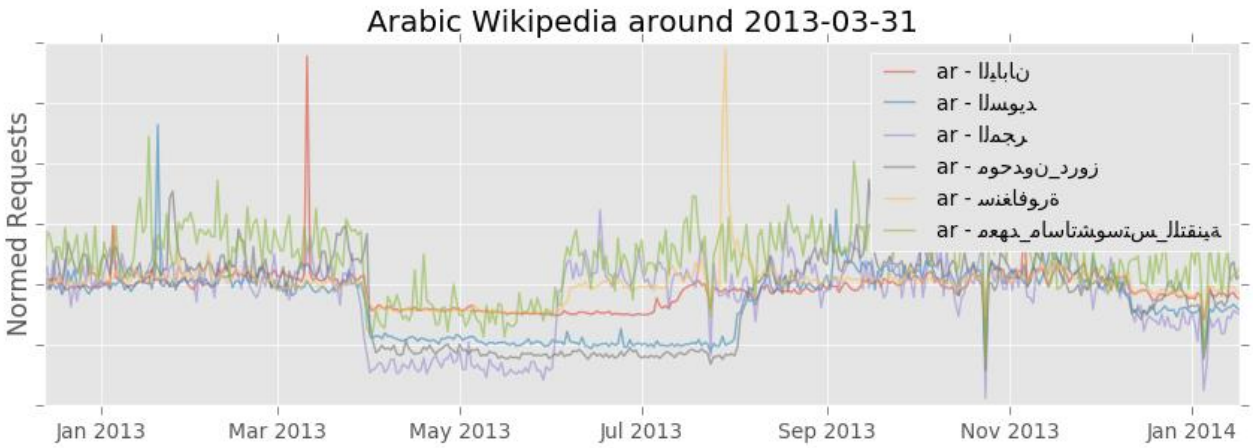
- 2011-12-18 through 2011-12-24
- 2013-03-28 through 2013-05-01
- 2013-07-21 through 2013-07-24
- 2013-10-21 through 2013-10-24
- 2014-01-02 through 2014-01-06
- 2014-12-01 through 2014-12-04
- 2014-12-19 through 2014-12-25
- 2015-05-05 through 2015-05-09
- 2015-11-15 through 2015-11-19



During the spring of 2013, a particular type of anomaly recurred across many thousands of articles across at least 34 different Wikipedia projects. The anomaly itself consisted of a group of seemingly unrelated articles from a single project simultaneously losing a significant amount of traffic. After some amount of time, usually between four and ten weeks, each article would return to normal levels. Initially thought to be true anomalies, anomalies following this pattern proved to be so widespread that it was deemed highly unlikely they were all caused by client-side events.

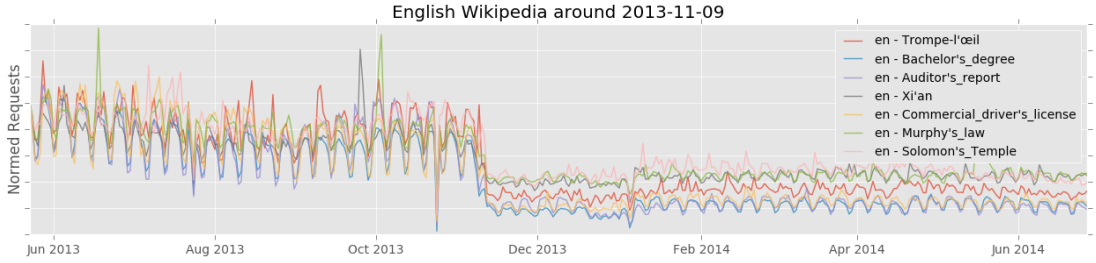
We identified anomalies that fit this pattern for articles within the following projects: bg, bs, cs, da, de, el, en, es, fi, fr, he, hr, hu, id, it, ja, lt, lv, ms, nl, no, pl, pt, ro, ru, sh, sk, sl, sv, th, tr, uk, vi, and zh. The start times of these anomalies are mostly concentrated in the period from March 28 through May 1, 2013, and this period of time was therefore omitted from further analysis. Examples of these anomalies from Spanish and Arabic Wikipedia follow:



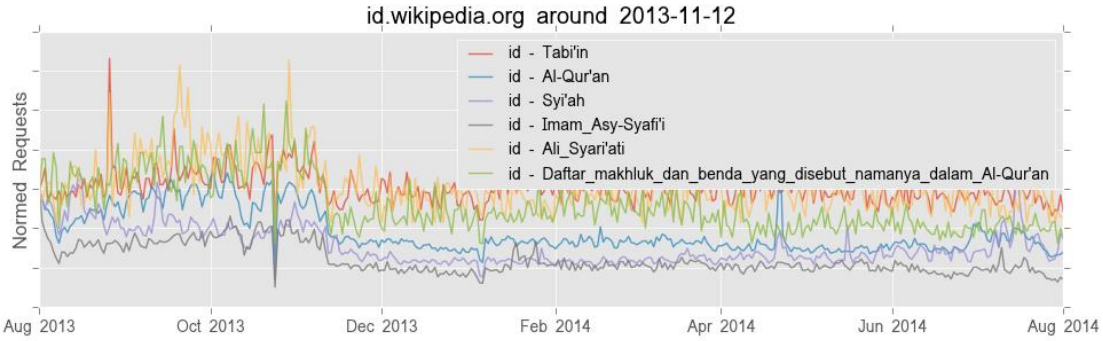


A shorter but similar event appears to have taken place in mid-June 2013. This period was not omitted from analysis, but events from this period were given more scrutiny.

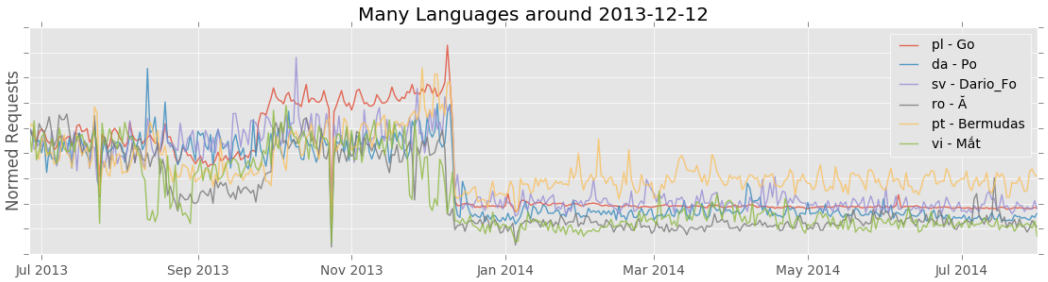
A number of dates saw widespread anomalies there were harder to classify, but did not appear related to censorship. Anomalies that occurred on or around these dates were considered less likely to be due to censorship, and more likely to be part of the underlying event causing widespread anomalies. Until the manual review of anomalies for many projects had taken place and the cross-project nature of these events were uncovered, many of these events were considered possible blocking events. For instance, starting around November 9, 2013, articles across a wide number of projects saw distinct drops in traffic. Interestingly, and so far inexplicably, most (though not all) of these articles contained apostrophes in their titles:



English Wikipedia saw highly significant dropoffs like those illustrated above in at least 32 articles containing apostrophes in the title. One fact that often made this event appear related to censorship was that in many languages, the articles most likely to contain apostrophes in the title are often related to Islam:

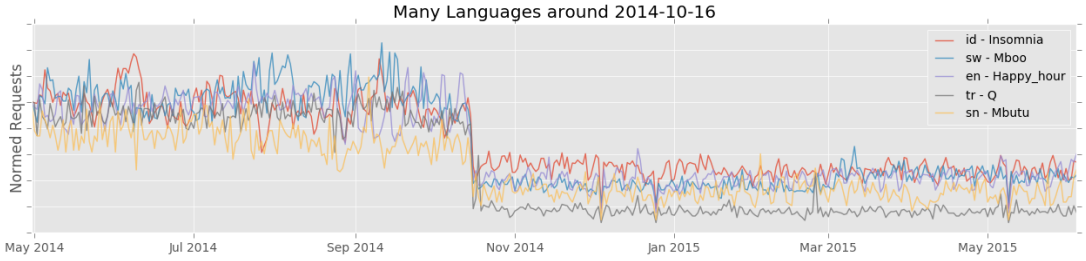


Beginning around December 9, 2013 and peaking around December 12, 2013, more than 175 articles from 27 different projects saw significant and sustained decreases in traffic. Many of the articles were for single letters ("A," "P," "Ă," "Γ"), and at least six projects saw significant drops for the article "Go":



While there were definitely patterns in the articles that saw significant drops, it proved difficult to attribute the pattern to any single cause.

October 16, 2014 was the peak of another set of anomalies. While this event took place for at least 14 Wikipedia projects, Japanese and Indonesian Wikipedias were the projects with the most affected articles. No clear relationship or patterns existed among the articles:



Appendix E: Article Analysis Methods In-Depth

As a method of inferring potential censorship motivation and providing context, we set out to identify articles that had been blocked in the past. In this idea's original iteration, we planned on detecting downward anomalies in the number of requests per day from each country to each Wikipedia article for as far back as the historical data would allow. The hypothesis was that fast, dramatic drops in the number of requests for an article do not occur as organic traffic patterns; rather, they must be the product of events that serve to move requests elsewhere or terminate the requests altogether. Our hope was that by locating these drops in traffic, we would have a heuristic for locating likely censorship events.

We started our analysis by creating a full list of all articles across all languages that could potentially be tested for anomalies. We first assembled a list of all Wikipedia projects, which resulted in 292 distinct projects.[179] For each of these projects, we downloaded the publicly available "Base per-page data" dumps from April 7, 2016.[180] These dumps were then inserted into a database. This process resulted in a dataset of 39,208,980 articles. There are currently 249 ISO-3166-1 country codes.[181] If we were to check every article from every one of these countries, we would need to analyze almost 10 billion time series. Even analyzing at a speed of 100 time series per second, it would take more than three years of computation time to check all article-country pairs. For this reason, combined with the presumption that there are far fewer articles that have been blocked than articles that have not, we chose to limit our analysis to a smaller set of articles.

There were a number of methods we could have used to generate the list of articles to be analyzed. The most obvious of these methods is a random sample, but given our assumption that the set of censored articles is much smaller than the set of uncensored articles, we wanted a way to oversample the set of likely censored articles. One method would be the manual curation of a list of articles deemed more likely to see censorship. This manual curation exercise is something the Berkman Klein Center is intimately familiar with. In the past, through the OpenNet Initiative (ONI) and related projects, the Berkman Klein Center has spent months collecting and categorizing such lists. Like this project, lists for these prior projects were meant to cover dozens of countries. To accomplish this, we primarily utilized a network of on-the-ground experts to develop country-specific lists. This was a large and slow endeavor. We therefore knew that the manual assembly of a sizable corpus of articles across all of Wikipedia's projects could become an intensive project in itself. We wished to avoid the large time and effort costs of our previous methods while still utilizing some of this previous work.

The lists of URLs we crafted for ONI were largely irrelevant to this project, but the categories of content that more often saw censorship were highly relevant. ONI categorized content into four

[179] "List of Wikipedias," *Wikimedia Foundation Labs*, http://wikistats.wmflabs.org/display.php?t=wp.
[180] For example, http://dumps.wikimedia.your.org/zhwiki/20160601/zhwiki-20160601-page.sql.gz.
[181] "ISO 3166-1," *Wikipedia*, https://en.wikipedia.org/wiki/ISO_3166-1.

broad categories: "Political," "Social," "Conflict/Security," and "Internet Tools". Internally, each of these categories contained a number of more specific topics (a total of 37 topics in all). For example, content in the Political category related to one of twelve topics (freedom of expression, women's rights, political reform, etc.) and the Social category contained nine topics (family planning, pornography, gambling, drugs, etc.). We had tweaked our taxonomy since the conclusion of ONI to contain 40 topics within the same four categories. Past ONI research indicated that each of the four broader categories saw pervasive censorship in at least one country.[182] Research on censored Wikipedia articles in the past has uncovered topics that were broadly similar to our own.[183] For these reasons, we decided that any method for constructing a set of articles must generate a set such that all four of our broad categories and a majority of the 40 topics within our taxonomy received analysis.

With the constraints that our sampled set must contain articles from a number of specific topics, that it must touch on most, if not all, of Wikipedia's projects, and that it must not become a sizable project unto itself, we designed an article selection method. The method we chose is as follows: we manually collected a list of "seed" articles that we deemed more likely to see censorship actions as the basis of our article set; we then added to this set all translations of these seed articles; we finally added all articles that were directly linked to by those already in our set. The use of a seed set would limit the manual curation we would need to perform, collecting translations would broaden coverage across Wikipedia projects, and link traversal would dramatically increase the number of articles in the set while still retaining some level of semantic relatedness.

The translation and link traversal steps were straightforward to implement, but the creation of the seed list still needed some level of manual curation. To develop a list of articles likely to see censorship, we researched lists of Wikipedia articles that have seen censorship in the past. There are only a few such lists: a China-centric list that GreatFire.org maintains and checks regularly for censorship,[184] a small Persian-centric list developed by Small Media,[185] and a larger Persian-centric list developed as part of a research project at the University of Pennsylvania.[186] Unfortunately, the public availability of the Persian-centric lists was discovered after much of our analysis was already complete,[187] so of the existing lists, we only used GreatFire.org's as a contributor to our seed set.

---

[182] "Filtering Data," *OpenNet Initiative*, https://opennet.net/research/data.
[183] Nima Nazeri and Collin Anderson, "Citation Filtered: Iran's Censorship of Wikipedia," *Center for Global Communication Studies*, Nov 2013, http://www.global.asc.upenn.edu/fileLibrary/PDFs/CItation_Filtered_Wikipedia_Report_11_5_2013-2.pdf.
[184] "Censorship of Wikipedia Pages in China," *Greatfire.org*, https://en.greatfire.org/search/wikipedia-pages.
[185] "Closed society meets open information," *Small Media*, Apr 12, 2013, https://smallmedia.org.uk/old/content/81.html.
[186] Nima Nazeri and Collin Anderson, "Citation Filtered: Iran's Censorship of Wikipedia," *Center for Global Communication Studies*, Nov 2013, http://www.global.asc.upenn.edu/fileLibrary/PDFs/CItation_Filtered_Wikipedia_Report_11_5_2013-2.pdf.
[187] 334 articles on these lists ultimately found their way into our analysis set, though our coverage of Persian Wikipedia would have improved with the inclusion of these lists.

The GreatFire.org list contained coverage of many of the topics in our Internet Tools category ("Tor," "Facebook," "YouTube," "WeChat," etc.), but most of the other articles were related to matters specific to China ("Tibetan Buddhism," "Tiananmen Square," "Falun Gong," etc.). To increase the coverage of our target topics, we decided to add all articles included in English Wikipedia's list of controversial issues (and articles that included the "Controversial" template).[188] ("Controversial" in this sense refers to a high incidence of "edit wars," where edits to articles are repeatedly made and reverted.) The list of controversial issues included articles that covered 25 of our 40 targeted sensitive topics, with especially good coverage of social and political issues. Coupled with GreatFire.org list, this met our goal of covering our four broad categories and more than three-quarters of the more specific topics within our taxonomy. Past research has also found a correlation between controversial Wikipedia articles and state censorship efforts, at least for the Iranian case.[189] The topic coverage of our article set and this past research gave us some confidence that our selected sample of articles would contain a higher proportion of censored articles than a purely random sample.

After assembling and cleaning the combined GreatFire.org and controversial articles lists, our seed set contained 2,933 articles. Fetching all translations of these articles expanded our set to 44,611 articles. Using the MediaWiki API[190] and the mwclient library,[191] we then added to our set all Wikipedia articles to which these 44,611 directly link. This resulted in a set of 1,722,543 articles. These 1.7 million articles became our top priority for data collection and analysis. This set included articles from 286 distinct Wikipedia projects (out of the total 292), and 132 projects were represented by more than 10,000 articles.

We then attempted to locate data on the number of requests per day for each of these articles from every country. We had a particular date range in mind when looking for this data. If Wikipedia could identify from their own data the articles that were likely censored, they might be able to infer motivations of the censors. When Wikipedia moved to providing content solely by HTTPS in June 2015, censors likely lost the ability to discriminate between articles they wanted to censor and articles they did not. This meant the change to HTTPS-only content delivery likely caused Wikipedia to correspondingly lose their window into some of the intentions of the various censoring bodies. Because of this possibility, we chose to look most closely at article-specific censorship prior to the June 2015 change to HTTPS.

For privacy reasons, Wikimedia does not publicly release request data separated out by both article and country, so we were granted research access to one of Wikimedia's internal research databases

---

[188] "Wikipedia:List of controversial issues," *Wikipedia*, https://en.wikipedia.org/wiki/Wikipedia:List_of_controversial_issues.
[189] Nima Nazeri and Collin Anderson, "Citation Filtered: Iran's Censorship of Wikipedia," *Center for Global Communication Studies*, Nov 2013, http://www.global.asc.upenn.edu/fileLibrary/PDFs/CItation_Filtered_Wikipedia_Report_11_5_2013-2.pdf.
[190] "API:Main page," *MediaWiki*, https://www.mediawiki.org/wiki/API:Main_page.
[191] "Mwclient," *Github*, https://github.com/mwclient/mwclient.

under a non-disclosure agreement. Upon entering the database, we discovered that data of this kind was only available from May 10, 2015 onward. We hypothesized that the transition Wikipedia made to HTTPS-only for all its projects in mid-June 2015 would eliminate much of the article-level censorship, and that therefore, we would only have been left with useful historical data from mid-May 2015 to mid-June 2015. As our workflow was designed primarily to locate the beginning of censorship events, it was deemed that the effort and time required to look for the beginning of censorship events in the four week window from mid-May to mid-June 2015 would not have been effort well spent.

The time and effort required to extract data from Wikimedia's internal research database played a part in this calculation. A query to extract one year's worth of daily request counts for a single article took approximately ten minutes. This query time would have been less of an issue if we had extracted data for many articles at once, as querying for multiple articles in the same query did not significantly increase the response time, but we then would have faced the issue of managing and querying against a large quantity of exported data on infrastructure we had little ability to control. We did explore using scratch space within the same database infrastructure to manage this exported data, but the database software introduced significant time overhead in even simple queries against this relatively small dataset that would have created a large bottleneck in our analysis pipeline.

Instead of investing significant work for four weeks of data, we shifted our focus to sources with more historical coverage. We identified four publicly available sources of historical article-level data: the Wikimedia Pageview API, the http://stats.grok.se site, the Wikipedia "pagecounts-raw" dumps, and the Wikipedia "pagecounts-ez" dumps. Because all these sources are meant for public consumption, much of the granularity had been removed prior to publication to protect user privacy.[192] That is most notable for this project because requests were no longer broken out by both article and the geographic location of the request. This meant that any analysis performed on the data could not connect censorship events directly to the countries within which the censorship was likely taking place. This was an unfortunate concession that needed to be made in order for our analysis to continue. With that decision made, we continued to evaluate the utility of the various data sources.

The Wikimedia Pageview API was quickly eliminated because its historical data starts July 1, 2015, which is after the transition to HTTPS. As we knew we only wanted data on 1.7 million of approxiately 40 million articles, and that we wanted this data daily rather than hourly, we chose to use the http://stats.grok.se API. This was convenient because we wouldn't need to download data for articles we were not interested in, and stats.grok.se had already aggregated requests by day whereas the dumps provided hourly data. stats.grok.se also provided the benefit of historical data back to December 2007.

[192] An example of Wikimedia's privacy efforts can be seen here: https://wikitech.wikimedia.org/wiki/Analytics/Data/Pageview_hourly/Sanitization.

Unfortunately, it quickly became clear that fetching data from the stats.grok.se API at the volume we needed was much too slow. We turned our attention to the pagecount data dumps. It was determined that the "pagecounts-raw" data was larger than we could handle with our storage infrastructure, which left the "pagecounts-ez" dumps as the most suitable source of data. This data came with a cost: instead of historical data back to December 2007, which both stats.grok.se and the "pagecounts-raw" dumps provided, the "pagecounts-ez" data only existed from November 2011 onward. This meant that we would not have data on censorship events that might have occurred in the period between December 2007 and November 2011. Again, this concession was necessary for our analysis to continue.

We downloaded the "pagecounts-ez" data (through the your.org mirror,[193] which provided substantially greater speeds), pulled out data on the 1.7 million articles we had selected earlier, and reaggreated the number of requests by day rather than by hour.

Now that we had data to analyze, we began the anomaly detection process. The intent of this analysis was to locate possible article censorship events. We used anomaly detection not as a statistical tool, but rather as a search heuristic to find events worth investigating. The core of this process was the Robust Principal Component Analysis (RPCA) anomaly detection algorithm.[194] This algorithm was chosen because it is moderately fast, works well across various types of time series, provides feedback on how anomalous each data point is, has open source implementations in multiple languages,[195] and was being used in production at Netflix.[196] When initiating this project, we designed our analysis pipeline around the constraint that much of the data could not leave Wikimedia's servers, and that therefore much of the analysis itself would need to take place on Wikimedia's servers. We therefore chose a pipeline architecture that was easy to deploy in an environment over which we had little control. Many of the requirements of this pipeline were fulfilled by Mozilla's Heka project.[197] Heka was attractive because it is written in Go, which meant we could simply compile a dependency-less binary, copy it to Wikimedia's servers, and feed it Wikimedia's data. For this to happen, we needed to have a version of the RPCA algorithm we could compile into Heka. That necessitated porting the RPCA algorithm to Go, which we did.[198] We further customized the Heka pipeline by adding the ability to aggregate data at different timescales. We also created two modules: one that grouped consecutive anomalous measurements together into multi-day anomalous events and computed a score for each event based on an aggregation of each constituent day's anomalousness and the total duration of the anomaly, and one to output anomalies

[193] http://dumps.wikimedia.your.org/
[194] Candes, et al., "Robust Principal Component Analysis?," http://statweb.stanford.edu/~candes/papers/RobustPCA.pdf.
[195] https://github.com/Netflix/Surus
[196] Chris Colburn, "RAD - Outlier Detection on Big Data," *Netflix Tech Blog*, http://techblog.netflix.com/2015/02/rad-outlier-detection-on-big-data.html.
[197] Heka's homepage: http://hekad.readthedocs.io/.
[198] The code is available here: https://github.com/berkmancenter/rpca.

to a CSV file for further analysis. Collectively, these new features were included in two open-sourced Heka plugins.[199] [200]

With everything in place, we ran each of our 1.7 million time series through the anomaly detection pipeline and output all the resulting anomalous events to an Elasticsearch index. We ultimately ended up with about 92.4 million anomalous events, 18 million of which were scored less than zero, indicating that the observed number of requests was lower than what the RPCA algorithm would have expected given the article's history. We then began looking through the events. Our first observation was that many of the most extreme anomalous events were short events that occurred at the same time across all or almost all articles. These events are outlined in Appendix D. We surmised that these were likely data collection issues on Wikimedia's side rather than actual Wikipedia-wide request drop offs. While Wikimedia does document many of their data collection issues,[201] we were unable to locate issue documentation going back far enough in time to confirm our suspicions. These events were excluded from further analysis.

Once those events were excluded, we generated graphs of the 500 anomalies with the lowest scores per Wikipedia project. We observed that many of the top anomalies were for articles with very little request traffic, and it appeared as though the algorithm was therefore picking up any traffic to these articles as anomalous. These results were undesirable, so for all articles that contained at least one anomaly with a score less than zero, we computed the median number of requests per day. We then regenerated our graphs for the 500 lowest scoring anomalies per project, but only considered articles that had a median of ten or more requests per day. 37 projects contained zero articles that fit that description, 118 projects contained more than zero but less than 500, and 137 projects contained 500 or more. Altogether we graphed 95,603 anomalies. To ensure we did not miss any notable events in large projects, we also graphed the 5000 anomalies with the lowest scores across all languages with more than 100 median requests per day. In total we selected and graphed 100,603 anomalies.

We then manually reviewed these anomaly graphs. This process was meant to familiarize us with the kinds and categories of anomalies we might be detecting, and how the shapes of these anomalies might indicate different phenomena. We investigated many anomalies for each of the various types of shapes that we saw. We could only find strong evidence of the processes backing two kinds of anomalies: national holidays and article editing events internal to Wikipedia. National holidays often see articles drop off quite dramatically (but never instantly), and they often recover just as quickly. In graphical form, they often look like sharp letter V's. Article editing events like moves, deletions, and redirects were behind a large number of detected anomalies in the number of requests. The detected changes could be negative or positive depending on the nature of the change. When searching

[199] Heka anomaly plugin: https://github.com/berkmancenter/hekaanom.
[200] Heka CSV encoder: https://github.com/berkmancenter/csvencoder.
[201] "Analytics/Data/Pageview hourly," *WikiTech*, https://wikitech.wikimedia.org/wiki/Analytics/Data/Pageview_hourly#Changes_and_known_problems_since_2015-06-16.

through the detected events for anomalies that might constitute blocking, we had to consider the fact that editing events could look very similar to our hypothesized censorship events.

We hypothesized that a censorship event would look like a relatively stable number of requests followed by an instant drop that stays stable for some period of time. If the article were to become unblocked, we also hypothesized that the number of requests would more slowly increase back to somewhere near pre-censorship levels. While reviewing the graphs, we paid special attention to graphs with this shape. We pulled many of these out for further investigation, along with other anomalies that did not look like they could be organic traffic patterns. We often chose not to select anomalies that occurred around national holidays and had the deep V shape we had previously associated with holidays. Altogether, we pulled out 1,288 anomalies. We then grouped anomalies that started on the same date and had similar shapes. For each of the groups of anomalies and the anomalies that did not fall into groups, we organized them into either high or low investigatory priority based on a number of judgements. First, we attempted to find anomalies that indicated a fast and severe drop in traffic that lasted for more than a single day. We also hypothesized that where traffic dropped completely to zero, these were likely data collection errors, and so we focused on anomalies where traffic dropped significantly but not completely. We attempted to find anomalies that occurred together in time, as we believed censors are likely to block more than a single article at a time. For anomalies that occurred together in time, we wanted to make sure they were limited to only a small number of languages and did not occur widely across all of Wikipedia's projects because we believed censors were more likely to target only the languages spoken within their country and that wider spread anomalies were more likely to be data collection issues. For the anomalies that remained, we considered them more likely to be related to censorship if the articles were thematically related, as synchronized anomalies for apparently unrelated articles could be caused by a larger number of processes (bots ceasing operations, the blocking of high-volume requesters, network outages, etc.). Once this review process was complete, we were left with 441 high priority anomalies and 847 lower priority anomalies. We then further investigated the high priority anomalies and only looked at lower priority anomalies that related to our countries of interest.

For the high priority anomalies, further investigation consisted of a number of steps. First, for each anomaly or group of anomalies, we queried and graphed the twenty anomalies with the largest drop offs starting on the same day for the same language or languages. We then reviewed these new anomalies for any that might fit the pattern of the high priority anomaly. Any that fit were added to the anomaly group. Then, for each anomaly or group of anomalies, we spot checked the histories of a number of articles to identify any events that might have been the cause of the anomalies. We checked both the edit history page of the article itself as well as the public log of the appropriate language.[202] If an editing event that could cause a significant decrease in traffic (page move, deletion) took place at the same time as a detected anomaly, we assumed this event was the cause of the anomaly and ruled it out as a potential censorship event. This step eliminated many of the top

[202] For example, for English we used https://en.wikipedia.org/wiki/Special:Log/all.

anomalies. Our next step was to confirm that the anomalies were not taking place on national holidays. For this step, we relied heavily on timeandate.com, which maintains a list of dates of historical holidays for many countries around the world.[203] For the anomalies that remained, if their article titles were not in English, we translated them using Google Translate[204] as a first pass and then located the English translation of the article in question for those translations that we considered suspect. Once we had translations, we took special care investigating anomalies in articles that looked to be thematically similar. The result of these investigations constitute the bulk of the article-level results presented above. As the process above illustrates, we chose to be conservative rather than comprehensive in our results.

[203] *Timeanddate.com*, http://www.timeanddate.com/holidays/.
[204] *Google Translate*, https://translate.google.com/.