Automatically Generating a Large, Culture-Specific Blocklist for China

Austin Hounsel
Princeton University

Prateek Mittal
Princeton University

Nick Feamster
Princeton University

Abstract

Internet censorship measurements rely on lists of websites to be tested, or "block lists" that are curated by third parties. Unfortunately, many of these lists are not public, and those that are tend to focus on a small group of topics, leaving other types of sites and services untested. To increase and diversify the set of sites on existing block lists, we use natural language processing and search engines to automatically discover a much wider range of websites that are censored in China. Using these techniques, we create a list of 757 websites outside the Alexa Top 1,000 that cover Chinese politics, minority human rights organizations, and oppressed religions. Importantly, *none of the sites we discover are present on the current largest block list*. The list that we develop not only vastly expands the set of sites that current Internet measurement tools can test, but it also deepens our understanding of the nature of content that is censored in China. We have released both this new block list and the code for generating it.

1 Introduction

Internet censorship is pervasive in China. Topics ranging from political dissent and religious assembly to privacy-enhancing technologies are known to be censored. Naturally, the list of sites that are filtered are not public [9]. When performing measurements of Internet filtering, the inability to know what sites are blocked creates a circular problem of discovering the sites to measure in the first place. To do so, various third parties currently curate lists of websites that are known to be censored, or "block lists". These lists are used to both understand what content is censored in China and how that censorship is implemented.

Instead of curating a block list by hand, Darer et al. proposed a system called FilteredWeb that automatically discovers webpages that are censored in China [7]. Their approach is summarized in the following steps. First, keywords are extracted from webpages on the Citizen Lab block list, a small, hand-curated list. These are English words that are ranked through TF-IDF, a technique which we describe in Section 3.2. Then, each keyword is used as a query for a search engine, such as Bing. The intuition is that censored webpages contain similar keywords. Finally, each webpage that appears in the search results is tested for DNS manipulation in China. These tests are performed by sending DNS queries to IP addresses in China that don't belong to DNS servers. If a DNS response is received, then, it is inferred that the request was intercepted in China, which implies that the website is censored. Each webpage that is censored is fed back to the beginning of the system. FilteredWeb discovered 1,355 censored domains, 759 of which are outside the Alexa Top 1,000.

In this paper, we build upon the approach of FilteredWeb in the following ways. First, *we extract content-rich phrases for search queries*. In contrast, FilteredWeb only uses single words for search queries. These phrases provide greater context regarding the subject of censored webpages, which enables us to find websites that are very closely related to each other. For example, consider the phrase 中国侵犯人权 (Chinese human rights violation). When we perform the searches `Chinese`, `human`, `rights` , and `violation` independently, we mainly get websites for Western media outlets, many of which are known to be censored in China. By contrast, if we search for [`Chinese human rights violation`] as a single phrase, then we discover a significant number of websites related to Chinese culture, such as homepages for activist groups in China and Taiwan. Identifying and extracting such key phrases is a non-trivial task, as we discuss later.

Second, we use natural language processing to parse Chinese text when adding to the blocklist. In contrast, FilteredWeb only extracts English words that appear on a webpage. As such, *any website that is written in simplified Chinese is ignored*, neglecting a significant portion of censored sites. For example, there are many censored websites and blogs that cover Chinese news and culture, and many of them only contain Chinese text. As such, to discover region-specific, censored websites, such a system should be able to parse Chinese text. Because Chinese is typically written without spaces separating words, this requires the use of natural-language processing tools.

Third, we make our block list public, in contrast to previous work. The authors of FilteredWeb made their block list available to us for validation; we have published our block list so others can build on it [12].

In summary, we built and now maintain a large, public, culture-specific list of websites that are censored in China. These websites cover topics such as political dissent, historical events pertaining to the Chinese Communist Party, Tibetan rights, religious freedom, and more. Furthermore, because many of these website are written from the perspective of Chinese nationals and expatriates, we are able to get first-hand accounts of Chinese culture that are not present in other block lists. This new resource can help researchers who are interested in studying Chinese censorship from the perspective of marginalized groups that most affected by it.

In this paper, we make the following contributions:

- We build upon the approach of FilteredWeb to discover censored websites in China that specifically pertain to its culture. We do so by extracting potentially sensitive Chinese phrases from censored webpages and using them as search terms to find related websites.
- We build a list of 757 censored domains in China, which is 8.5× larger than the standard list for censorship measurements [5]. Furthermore, *none of these websites are on the largest block list available* [7].
- We perform a qualitative analysis of our block list to showcase its advantages over previous work.

The rest of the paper proceeds as follows. First, we describe our approach to building a large, culture-specific block list for China. This includes an in-depth analysis of the advantages of our approach over previous work. Then, we describe two large-scale evaluations that we performed. Each of these evaluations produced qualitatively different results due to different configurations of our system. Finally, we conclude with a discussion of how the block list we built could be used by researchers. We also briefly explore directions for future work.

2 Related work

Existing block lists have several limitations. First, some of these block lists have been curated by organizations that study Internet censorship, but these lists are now outdated [4, 18]. Other lists have been automatically generated by systems similar to ours, but they are not publicly available at the time of publication [7, 8, 25]. Furthermore, these systems do not focus on blocked websites that are particular to Chinese culture. There are also systems that create block lists through crowd-sourcing, but are unable to automatically detect newly censored websites [10, 11]. Finally, the Citizen Lab block list focuses on popular websites–such as social media, Western media outlets, and VPN providers–but not websites pertaining to Chinese culture [5].

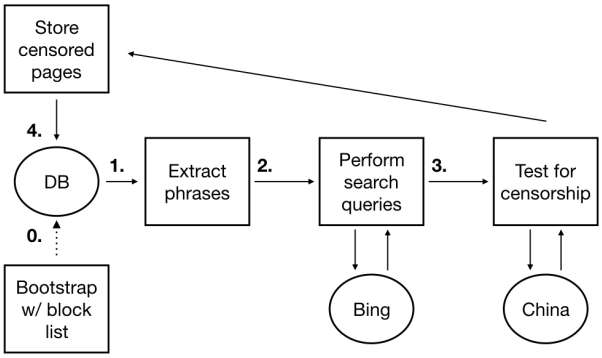

Figure 1: *Our approach for discovering censored websites.*

Other systems use block lists to determine *how* censorship works, but they do not create more block lists. For example, Pearce et al. proposed a system that uses block lists to measure how DNS manipulation works on a global scale by combining DNS queries with AS data, HTTPS certificates, and more [20]. Pearce et al. also built Augur, a system that utilizes TCP/IP side channels between a host and a potentially censored website to determine whether or not they can communicate with each other [19]. Furthermore, Burnett et al. proposed Encore, a system that utilizes cross-origin requests to potentially censored websites to *passively* measure how censorship works from many vantage points [2]. Lastly, several platforms have been proposed that crowd-source censorship measurements from users by having them install custom software on their devices [14, 17, 23].

3 Approach

Figure 1 summarizes our approach to finding censored websites. For the most part, our approach is similar to that of FilteredWeb [7]. We start by bootstrapping a list of webpages that are known to be censored, such as the Citizen Lab block list [5]. Then, we extract Chinese and English phrases that characterize these webpages. Then, we use these phrases as search queries to find related webpages that might also be censored. Finally, we test each search result for DNS manipulation in China and feed the censored webpages back to the beginning of the system. The rest of this section details the new capabilities of our approach beyond the state of the art.

3.1 Extracting multi-word phrases

An n-gram is a building block for natural-language processing that represents a sequence of "n" units of text, e.g. words. For example, if we were to compute all of the bigrams of words in the English phrase [`Chinese human rights violation`], we'd get `Chinese human`, `human rights`, and `rights violation`. Similarly, if we were to compute all the trigrams of words, we'd get `Chinese human rights` and `human rights violation`. On the other hand, if we simply extract unigrams from the sentence, we'd be left with `Chinese`, `human`, `rights`, and `violation`. As such, we believe that by combining unigrams into bigrams and trigrams, we're able to learn more about the subject of a webpage.

Unfortunately, computing the n-grams of words in Chinese text is difficult because it is typically written without spaces, and there is no clear indication of where characters should be separated. Computing such boundaries in natural-language processing tasks is often probabilistic, and depending on where characters are separated, one can arrive at very different meanings for a given phrase [26]. For example, consider the text 天花. This is considered a Chinese unigram that translates to "smallpox" in English, but the character 天 on its own translates to "sky", and the character 花 on its own translates to "flower". Given that we do not have domain-specific knowledge of the webpages that we're analyzing, we consider a unigram to be whatever the segmenter software in the Stanford CoreNLP library considers to be a unigram, even if a Chinese unigram translates to multiple words in English [3, 28].

We believe that using Chinese bigrams and trigrams for search queries instead of individual English words is more effective for discovering censored websites in China. For example, we know that websites that express collective dissent are considered sensitive by the Chinese government [15]. If we used the words `destroy`, `the`, `Communist`, or `Party` individually as search terms, we might not get websites that express collective dissent because the words have been taken out of context. On the other hand, the phrase [`destroy the Communist Party`] as a whole expresses collective dissent, which might enable us to find many censored domains when used as a search term. We evaluate the effectiveness of such phrases in Section 5.3.

3.2 Ranking phrases

To "rank" the phrases on a censored webpage, we use term-frequency/inverse document-frequency (TF-IDF). TF-IDF is a natural language technique that allows us to determine which phrases best characterize a given webpage [22]. It can be thought to work in three steps. First, we compute the term-frequency for each phrase on a webpage, which means that we count the frequency of each bigram and trigram. Then, we compute the document-frequency for each phrase on a webpage, which entails searching a Chinese corpus for the frequency of a given phrase across all documents in the corpus [21]. Finally, we multiply the term frequency by the inverse of the document frequency. The resulting score gives us an idea of how important a given Chinese phrase is to a webpage.

Using this method, phrases like 1989年民主运动 [`1989 democracy movement`] and 天安门广场示威 [`Tienanmen Square Demonstrations`] might rank highly on a website about Chinese political protests. We would then use these phrases as search terms in order to find related websites that might also be censored. If we find a lot of censored URLs as a result, then we can infer that the topic covered by that phrase is considered sensitive in China.

3.3 Parsing Chinese text

Before we can compute TF-IDF for a given webpage, we need to tokenize the text. Doing so is simple enough for English because each word is separated by a space. For languages such as Chinese, however, all of the words in a sentence are concatenated, without any spaces between them. As such, we need to apply natural-language processing techniques to perform fine-grained analysis of the text on a webpage.

To do so, we make use of Stanford CoreNLP, a set of natural language processing tools that operate on text for English, Arabic, Chinese, French, German, and Spanish [6]. For Chinese webpages, it allows us split a sentence into a sequence of unigrams, each of which may represent one word or even multiple words. By combining neighboring unigrams, then, we can extract key phrases from a webpage that describe its content.

As previously mentioned, although FilteredWeb is concerned with finding webpages that are censored in China, it is only able to parse text for the ISO basic Latin alphabet [7]. By being able to parse Chinese text as well as ISO Latin text, then, we are able to cover many more webpages and extract regional information that may explain why a given webpage is censored. In future work, we could make use of more intricate tools from Stanford CoreNLP likes its part-of-speech tagger to identify phrases that convey relevant information.

4 Evaluation

We performed a large-scale evaluation of our approach to discover region-specific websites that are censored in China. We began by seeding with the Citizen Lab block list, which is the most widely used list by censorship researchers. The list contains 220 webpages that either are blocked or have been alleged to be blocked in the past. Since we only extract phrases from webpages that are

currently censored, we began by testing each webpage on the block list for censorship. This left us with 108 unique webpages and 85 domains.

From November 11th, 2017 to January 9th, 2018, we used Bing's Search API to search for websites related to known censored websites [16]. According to the API, each call can return at most 50 search results. Since it would be expensive to perform multiple API calls for each search term, we limited ourselves to one API call, i.e. 50 search results, per search term. For each website in the search results, we tested for censorship by sending a DNS request to a set of controlled IP addresses in China that don't belong to DNS servers. As with FilteredWeb, if we received a DNS response, we inferred that the DNS request was intercepted, and thus the tested website is censored [7]. Finally, we performed two separate evaluations to measure the effectiveness of different phrase sizes for finding censored websites. In the first evaluation, we only extracted bigrams of words, and in the second one we only extracted trigrams.

We also configured the Bing API calls so that any URLs from Blogspot, Facebook, Twitter, YouTube, and Tumblr would be ignored. We did this for a couple of reasons. For one, these websites are widely known to be censored in China, so we would not be providing new information by having these websites or their subdomains in our result set. Furthermore, Tumblr and Blogspot assign a unique subdomain to each user's blog, and in some cases we'd get a dozen blogs from a single search query. In order to find culture-specific websites that are normally buried by the top 50 search results, then, we need to omit these blogs.

5 Results

We reached several key insights from performing our evaluations. First, we were able to discover hundreds of censored websites that are not present on existing block lists. Furthermore, we noticed that many websites on our block list receive very little traffic. Lastly, we found that by using politically-charged phrases as search terms, we were able to find a disproportionally large amount of censored websites. The rest of this section discusses the main findings in depth.

Phrase length	Domains	Domains w/o Alexa Top 1000
Unigrams	1029	765
Bigrams	970	655
Trigrams	975	629
Total	1756	**1125**

Table 1: *Total number of censored domains discovered.*

5.1 Existing blocklists are incomplete

By using natural-language processing on Chinese webpages, we were able to discover hundreds of censored websites that are not present on the Citizen Lab block list–the standard for censorship measurements–and FilteredWeb's block list [5, 7]. Furthermore, the set of websites that appeared the most in our search results is almost entirely different from that of FilteredWeb. Figure 2 breaks down this result when using both bigrams and trigrams as search queries. These websites seem to mainly cover Chinese human rights issues, news, and political content.

For instance, `hrw.org` exposes a number of humans rights issues in China, including the persecution of the Ugyhur minority group [13]. Additionally, the `twreporter.org` covers Taiwanese news, which may seem like an innocuous website to Westerners. However, the relationship between Taiwan and China is strained to the point that the Chinese government has attempted to silence news outlets that report on Taiwan as its own country [27]. As such, the censorship of these popular websites might reflect China's domestic and foreign policy.

Finally, only three of the top 25 domains that we discovered are in the top 50 domain list produced by FilteredWeb. Four of the top five domains that we discovered–`www.storm.mg`, `www.plurk.com`, `buzzorange.com`, and `www.hk01.com`–are non-Western media websites that cover Chinese news. These websites were also not present in the list of top 50 domains discovered by FilteredWeb. *Thus, by building on the approach of FilteredWeb, we were able to produce a qualitatively different block list.* We recommend putting these two block lists together to create a single block list that is both wide in scope and large in size.

5.2 China blocks many unpopular websites

Figure 3 shows the ranking of the websites we discovered on the Alexa Top 1,000,000. Notably, many of the websites we discovered are spread throughout the tail of the list, and some of the websites are not even on the list at all. *Given that the top 100,000 websites likely receive the vast majority of traffic on the Internet, we can infer that censors in China are not just interested in blocking "big-name", popular websites.* They are actively seeking out websites of *any size* that contain "sensitive" content. We also discovered a number of websites that fall outside the Alexa Top 1,000,000 altogether. Without the use of an automated systen that can discover censored websites, it's unlikely that the public would even be aware that these websites are blocked.

Furthermore, we were able to consistently discover *more* unpopular censored domains with our approach than FilteredWeb. For domains that rank between 200,000 and 300,000 on the Alexa Top 1,000,000, FilteredWeb was able to discover more domains, but in every other bin of

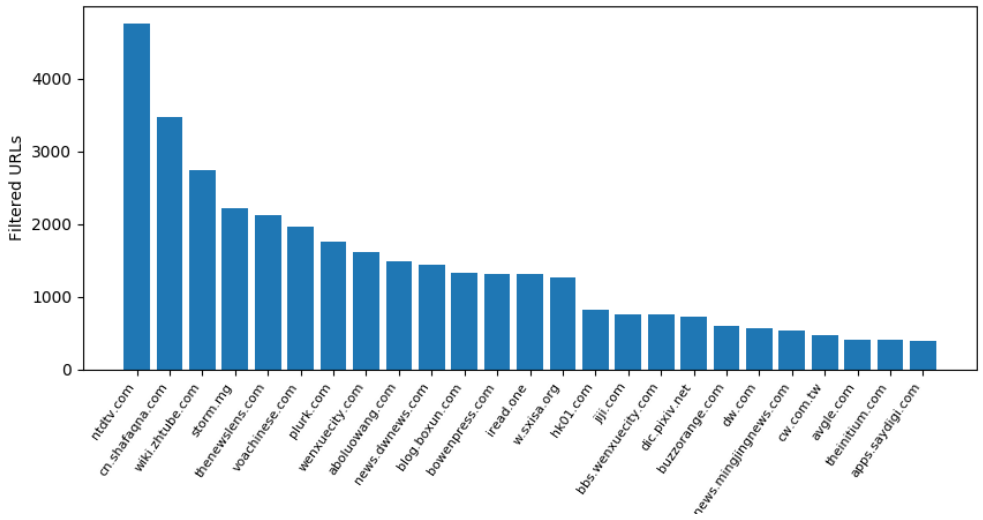

Figure 2: *Top censored domains by URL count.*

100,000 ranks we discovered more. Notably, between the ranks of 900,000 and 1,000,000, FilteredWeb did not discover *any* censored domains, and we discovered far more domains that fall outside of the Alexa Top 1,000,000 altogether. This suggests that the use of Chinese phrases as search queries allowed us to uncover a significant number of websites that would otherwise be unknown to censorship researchers.

Lastly, Table 1 shows the breakdown of how many censored websites we discovered. By using bigrams, we were able to discover 964 censored domains that are not on any publicly available block list, 655 of which are outside the Alexa Top 1000. Similarly, by using trigrams, we were able to discover 969 domains that are not on any publicly available block list, 629 of which are outside the Alexa Top 1000. In total, we found 1,405 domains, 757 of which are outside of the Alexa Top 1000 [1]. Each of these evaluations were performed with 1,000,000 unique URLs, consistent with the methodology of FilteredWeb [7]. Figure 4 shows how many censored domains we discovered as a function of unique URLs crawled for each evaluation.

5.3 Political phrases are highly effective

We also wanted to see if there is a correlation between the presence of certain phrases and whether or not a given website is censored. For example, if we make a search for 中国侵犯人权 (Chinese human rights violation) and find that a particular search result is censored, then we cannot be certain whether the presence of that phrase *caused* the website to be blocked. Even if we assume that a censor is manually combing through search engines to find sensitive websites, the website could have been blocked because it contained totally different content.

Nevertheless, there seems to be some correlation for both bigrams and trigrams. Table 3 shows a sample of bigrams that returned the most number of unique filtered domains from Bing. There are a couple of things to note. First, most of the bigrams refer to the Chinese Communist Party in some way. Phrases such as 中共的威胁 (Chinese Communists threaten), 江泽民胡锦涛 (Jiang Zemin Hu Jintao), and 中国共产党的治安 (Public security of the CPC) do not necessarily convey sensitive information, but they nonetheless refer to the government of China. On the other hand, some phrases clearly refer to political dissent, such as 官员呼吁 (Officials called on), 迫害活动 (Persecution activities), 非法拘留 (Illegal detention), and 宣称反共 (Declared anti-communist).

Table 4 shows a sample of trigrams that returned the most number of unique filtered domains from Bing. Phrases that stand out include 中国共产党的宗教政策 (The Chinese Communist Party's religious policy), 天安门广场示威 (Tiananmen Square demonstrations), 1989年民主运动 (1989 democracy movement), and 采取暴力镇压 (To take a violent crackdown). Interstingly, we also see that discussion of China's religious policy, the "New Tang Dynasty"–a religious radio broadcast located in the United States–, and European Union legislation may also be considered sensitive content. [24]. *Together, these results suggest that references to collective political dissent are highly likely to be censored.* This is consistent with the findings of King et al. [15].

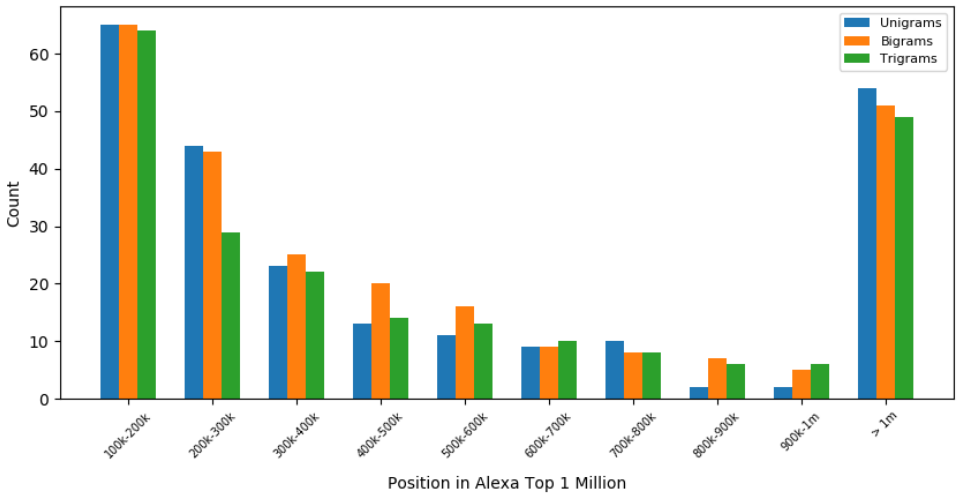

Figure 3: *Ranking of censored websites on Alexa Top 1 Million.*

Chinese	English	Censored domains
王歧山	Wang Qishan	74%
李洪志	Li Hongzhi	64%
郭伯雄	Guo Boxiong	62%
胡锦涛	Hu Jintao	56%
胡平	Hu Ping	54%
Morty	Morty	52%
命案	Homicide	52%
特首	Chief executive	52%
Vimeo	Vimeo	50%
中情局	CIA	50%

Table 2: *Sample of unigrams with significant blockrates*

6 Conclusion

We built a block list of 757 censored websites in China that are not present on the largest block list [7]. Furthermore, our list is 8.5× larger than the most widely used block list [5]. It contains human rights organizations, minority news outlets, religious blogs, political dissent groups, privacy-enhancing technology providers, and more. We've open-sourced our source code and block list on GitHub with the GPL license [12].

Automatically detecting which websites are censored in a given country is an open problem. Approaches like ours and that of FilteredWeb are steps in the right direction, but there is more work to be done. One way of improving our work would be to experiment with advanced natural-language processing techniques to identify better search terms. For example, we could try using Stanford CoreNLP's part-of-speech tagger to identify phrases that describe some action against the Chinese government. This approach might identify the following phrase: "Chinese citizens protest against the Communist Party on June 4th". We believe that using such culture-specific phrases as search terms would enable us to discover even more websites that are censored in China.

Acknowledgements

We thank the anonymous FOCI '18 reviewers for their helpful feedback. We also thank Chris Miller for help with setting up a machine that was used for data analysis. This work was supported by [...] and [...].

Chinese	English	Censored domains
中共 威胁	Chinese Communists threaten	50%
声明的 反共产主义	Declared anti-communist	44%
中国共 产党的公共安全	Public security of the CPC	42%
北京 清洁	Beijing clean-up	40%
江泽民 胡锦涛	Jiang Zemin Hu Jintao	40%
迫害 活动	Persecution activities	40%
官员 呼吁	Officials called on	36%
重的 公民	Heavy Citizen	36%
非法 拘留	Illegal detention	36%
不同 的民主	Different Democratic	34%

Table 3: *Sample of bigrams with significant block rates*

Chinese	English	Censored domains
北戴 河 会议	BEIDAIHE meeting	54%
中国 共产党 的宗教政策	The Chinese Communist Party's religious policy	42%
采取 暴力 镇压	To take a violent crackdown	38%
香港 政 治	Hong Kong Politics	34%
欧洲 议会 决议	European Parliament Resolution	32%
新 唐 王朝	New Tang Dynasty	32%
恐怖 事件 。	A terrorist event.	32%
天安 门广 场示威	Tiananmen Square Demonstrations	32%
敦促 美国 政府	Urging the US government	32%
1989 民主 运动	1989 democracy movement	30%

Table 4: *Sample of trigrams with significant blockrates*

References

[1] Alexa top sites. https://aws.amazon.com/de/alexa-top-sites/. Accessed:2017-12-13. (Cited on page 5.)

[2] Burnett, S., and Feamster, N. Encore: Lightweight measurement of web censorship with cross-origin requests. *ACM SIGCOMM Computer Communication Review 45*, 4 (2015), 653–667. (Cited on page 2.)

[3] Chang, P.-C., Galley, M., and Manning, C. D. Optimizing chinese word segmentation for machine translation performance. In *Proceedings of the third workshop on statistical machine translation* (2008), Association for Computational Linguistics, pp. 224–232. (Cited on page 3.)

[4] China Digital Times. China digital times: About us. https://chinadigitaltimes.net/about/. Accessed: 2018-05-25. (Cited on page 2.)

[5] Citizen lab block list. https://github.com/citizenlab/test-lists. Accessed: 2017-10-16. (Cited on pages 2, 4 and 6.)

[6] CoreNLP, S. A suite of core nlp tools. *URL http://nlp. stanford. edu/software/corenlp. shtml (Last accessed: 2013-09-06)* (2016). (Cited on page 3.)

[7] Darer, A., Farnan, O., and Wright, J. Filteredweb: A framework for the automated search-based discovery of blocked urls. In *Network Traffic Measurement and Analysis Conference (TMA), 2017* (2017), IEEE, pp. 1–9. (Cited on pages 1, 2, 3, 4, 5 and 6.)

[8] Darer, A., Farnan, O., and Wright, J. Automated discovery of internet censorship by web crawling. *arXiv preprint arXiv:1804.03056* (2018). (Cited on page 2.)

[9] Freedom House. Freedom on the net 2017: China. https://freedomhouse.org/report/freedom-net/2017/china. Accessed: 2018-02-04. (Cited on page 1.)

[10] Greatfire. http://greatfire.org. Accessed: 2017-11-15. (Cited on page 2.)

[11] Herdict. Herdict: About us. https://www.herdict.org/about. Accessed: 2018-05-25. (Cited on page 2.)

[12] Hounsel, A. Censorseeker source code and block lists. https://github.com/ahounsel/censor-seeker. Accessed: 2018-07-01. (Cited on pages 2 and 6.)

[13] Human rights watch: Uighurs. https://www.hrw.org/tag/uighurs. Accessed: 2018-1-15. (Cited on page 4.)

[14] ICLab. The information controls lab. https://iclab.org. Accessed: 2018-05-29. (Cited on page 2.)

[15] King, G., Pan, J., and Roberts, M. E. How censorship in china allows government criticism but silences collective expression. *American Political Science Review 107*, 2 (2013), 326–343. (Cited on pages 3 and 5.)

[16] Microsoft azure search api. https://azure.microsoft.com/en-us/services/cognitive-services/bing-web-search-api/. Accessed: 2017-11-15. (Cited on page 4.)

[17] OONI. Open observatory of network interference (ooni): About. https://ooni.torproject.org/about/. Accessed: 2017-10-07. (Cited on page 2.)

[18] Open Net Initiative. Internet filtering in china in 2004-2005: A case study. https://opennet.net/studies/china. Accessed: 2018-05-29. (Cited on page 2.)

[19] PEARCE, P., ENSAFI, R., LI, F., FEAMSTER, N., AND PAXSON, V. Augur: Internet-wide detection of connectivity disruptions. In *Security and Privacy (SP), 2017 IEEE Symposium on* (2017), IEEE, pp. 427–443. (Cited on page 2.)

[20] PEARCE, P., JONES, B., LI, F., ENSAFI, R., FEAMSTER, N., WEAVER, N., AND PAXSON, V. Global measurement of dns manipulation. In *Proceedings of the 26th USENIX Security Symposium (Security'17)* (2017). (Cited on page 2.)

[21] Phrasefinder.io. `http:/phrasefinder.io`. Accessed: 2017-11-15. (Cited on page 3.)

[22] RAMOS, J., ET AL. Using tf-idf to determine word relevance in document queries. In *Proceedings of the first instructional conference on machine learning* (2003), vol. 242, pp. 133–142. (Cited on page 3.)

[23] RAZAGHPANAH, A., LI, A., FILASTÒ, A., NITHYANAND, R., VERVERIS, V., SCOTT, W., AND GILL, P. Exploring the design space of longitudinal censorship measurement platforms. *arXiv preprint arXiv:1606.01979* (2016). (Cited on page 2.)

[24] Religious freedom and restriction in china. `https://berkleycenter.georgetown.edu/essays/religious-freedom-and-restriction-in-china`. Accessed: 2018-1-15. (Cited on page 5.)

[25] SFAKIANAKIS, A., ATHANASOPOULOS, E., AND IOANNIDIS, S. Censmon: A web censorship monitor. In *USENIX Workshop on Free and Open Communication on the Internet (FOCI)* (2011). (Cited on page 2.)

[26] STANFORD NATURAL LANGUAGE PROCESSING GROUP. Stanford word segmenter. `https://nlp.stanford.edu/software/segmenter.html`. Accessed: 2018-07-01. (Cited on page 3.)

[27] The absurd face of china's censorship. `http://wapo.st/2euQsN1`. Accessed: 2018-1-15. (Cited on page 4.)

[28] TSENG, H., CHANG, P., ANDREW, G., JURAFSKY, D., AND MANNING, C. A conditional random field word segmenter for sighan bakeoff 2005. In *Proceedings of the fourth SIGHAN workshop on Chinese language Processing* (2005). (Cited on page 3.)

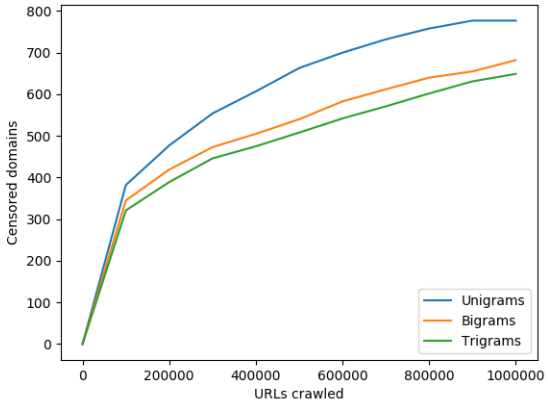

Figure 4: *Censored domains discovered over unique URLs crawled.*