Darwin's Theory Of Censorship: Analysing the Evolution of Censored Topics with Dynamic Topic Models

Asim Waheed*
a7waheed@uwaterloo.ca
University of Waterloo
Waterloo, ON, Canada

Sara Qunaibi*
squnaibi@uwaterloo.ca
University of Waterloo
Waterloo, ON, Canada

Diogo Barradas
diogo.barradas@uwaterloo.ca
University of Waterloo
Waterloo, ON, Canada

Zachary Weinberg
zackw@cmu.edu
Carnegie Mellon University
Pittsburgh, PA, USA

ABSTRACT

We present a statistical analysis of changes in the Internet censorship policy of the government of India from 2016 to 2020. Using longitudinal observations of censorship collected by the ICLab censorship measurement project [21], together with historical records of web page contents collected by the Internet Archive [17], we find that machine classification techniques can detect censors' reactions to events without prior knowledge of what those events are. However, gaps in ICLab's observations can cause the classifier to fail to detect censored topics, and gaps in the Internet Archive's records can cause it to misidentify them.

CCS CONCEPTS

• **General and reference** → **Measurement**; • **Social and professional topics** → **Censorship**.

KEYWORDS

censorship, internet measurement, topic modelling

ACM Reference Format:
Asim Waheed, Sara Qunaibi, Diogo Barradas, and Zachary Weinberg. 2022. Darwin's Theory Of Censorship: Analysing the Evolution of Censored Topics with Dynamic Topic Models. In *Proceedings of the 21st Workshop on Privacy in the Electronic Society (WPES '22), November 7, 2022, Los Angeles, CA, USA.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3559613.3563206

1 INTRODUCTION

Over the past twenty-five years, the Internet has increasingly become a prominent medium of mass communication, offering people around the world ready access to information and a venue to publish their opinions. As they do with television, radio, and newspapers, governments that view free expression as a threat to their legitimacy seek to control access to the Internet to quash dissent [3, 10, 29]. They use a variety of tactics to do this; the most obvious, and the best-studied, is to monitor network traffic for sensitive "keywords" and disrupt communications that contain them [6, 18, 19, 28].

Over the past decade, academic researchers in computer security have joined forces with political scientists and activists to develop systems that continuously monitor both *what* is censored by each repressive government and *how* the censorship is executed. Three of the most prominent such systems are Censored Planet [24], ICLab [21], and OONI [12]. In addition to their own publications, all three make raw data available to researchers. Previous analyses of this data suggest that blocked websites often fall into clusters that share a theme, or topic [7, 9, 21, 24, 26]. In some cases, these clusters can be identified with the aid of a URL classification service, such as FortiGuard [13]. However, sites in the "long tail" are frequently classified incorrectly, or not at all, and these are the sites whose censorship is most interesting to political analysts. (For example, FortiGuard is excellent at identifying online gambling sites but has nothing useful to say about many India-focused political blogs.)

In the absence of reliable manual classification for "long tail" websites, researchers have turned to machine classification. One frequently used technique is the Latent Dirichlet Allocation (LDA) algorithm, first introduced in 2003 by Blei et al. [5], which divides a text corpus into clusters that share a set of words with each other but not with the rest of the corpus. A human analyst then labels the clusters with subject headings based on the sets of words. Weinberg et al. [26] used LDA to assign topics to web pages suspected to be censored in 12 countries. Ramesh et al. [23] performed a similar study focusing on sites known to be blocked in Russia. Chen et al. [7] applied LDA to censored social media posts on Sina Weibo (China's alternative to Twitter) and Tanash et al. [25] classified censored content on Twitter within Turkey.

All of these efforts only looked at a single snapshot in time, but it is widely suspected that censorship policies are continuously updated, with current political controversy receiving more attention than historical events. To give a concrete example, Rambert et al. [22] compared Chinese keyword filtering in 2021 with a list published in 2014, finding that only 20% of the keywords from 2014 were still censored, and that new terms had been added. Similarly, the OONI organization [12] regularly publishes reports of specific

*Authors contributed equally.

WPES '22, November 7, 2022, Los Angeles, CA, USA

ACM ISBN 978-1-4503-9873-2/22/11…$15.00
https://doi.org/10.1145/3559613.3563206

sites, terms, or protocols that are newly blocked. It would be valuable to understand the evolution of censorship policy at the level of topic clusters as well as individual websites, and continuously, rather than by comparing snapshots.

In this paper, we present a pilot study in the use of dynamic topic modeling to capture the evolution of censored topics. We use Dynamic LDA [4], an extension to the LDA algorithm that divides a corpus into time slices and then uses LDA to classify the documents in each slice, while constraining how much the classification can change between any two adjacent slices. Dynamic LDA produces clusters, similar to LDA, but the membership of each cluster is a function of the time slice, providing insight into how the clusters evolve over time. Dynamic topic models have been used to model the evolution of the European parliament's political agenda [15] and to detect fraudulent schemes on cryptocurrency forums [20].

We summarize the contributions of this work as follows:

- We report the successful use of Dynamic LDA to characterize the evolution of Internet censorship policy from 2016 to 2020 by the Republic of India.
- We test the robustness of Dynamic LDA in the face of gaps in the underlying observations, and make recommendations for the granularity of future data collection efforts.

2 STUDY DESIGN

This study analyzes a subset of the ICLab data set. ICLab [21] has been collecting "longitudinal" observations of Internet censorship (i.e. repeated measurements at regular intervals) since 2017, from 62 different countries.

For this pilot study, we focus on censorship conducted by the Republic of India. Censorship in India is not as aggressive or as notorious as in, say, China, but Indian ISPs are legally required to be *capable* of blocking access to specific sites as directed by the government, and this capability has been regularly used [14, 29]. At the same time, ICLab has had no trouble collecting data in India. (Niaki et al. [21] observe that the countries that most aggressively censor the Internet are also the countries where they find it most difficult to measure censorship.) We particularly wanted to target a country where ICLab's data was as continuous in time as possible: as we will discuss further in Section 3.3, gaps in data collection seriously interfere with the accuracy of LDA-based topic modeling.

2.1 The Data Set

ICLab's "test list" for India comprises 6 012 unique URLs. Since March of 2016, ICLab has attempted to access these URLs 2 207 different times from vantage points in India. (Not every URL was tested every time.) ICLab records detailed information about each censorship event it observes. In this study, we conservatively consider a URL to be censored only if ICLab has flagged it as *overtly* censored by the Indian government or an Indian ISP. Overt censorship is when a client's connection is redirected to a "block page" which overtly states that access to the requested URL is forbidden. ICLab can also detect several types of "covert" censorship, where clients' connections are disrupted in a way that mimics an ordinary network fault. However, in the majority of cases, the events they detect this way are not *certain* to be censorship, so we excluded those

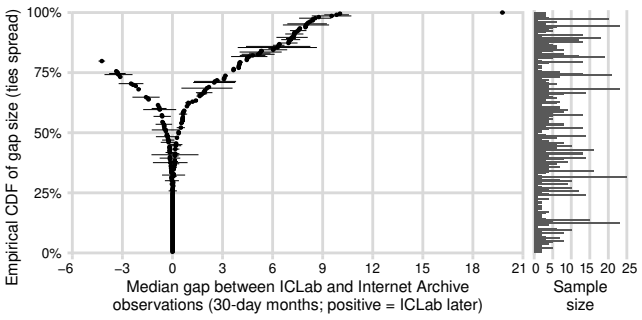


Figure 1: Empirical CDF of the time difference between Wayback Machine's snapshots and ICLab measurements.

events from this study. ICLab reports 677 of the 6 012 URLs were overtly censored at least once within the time period we analyzed.

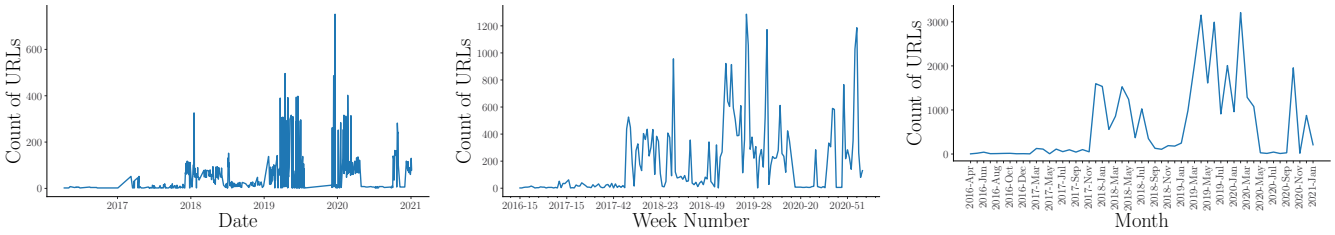
ICLab records the *apparent* text of each web page at the time of each observation. However, when censorship occurs, that text is either nonexistent or is supplied by the censor. For this study, what we require is the *uncensored* text of each web page at the time of each observation. We acquired uncensored text from the Internet Archive's "Wayback Machine" [17], which regularly collects snapshots of many websites from a location not subject to censorship. It has collected at least one snapshot for 603 of the 677 censored URLs. Due to resource limitations, we could only include 213 of these URLs (35%) in this pilot study.

Unfortunately, the Wayback Machine's snapshots are not always simultaneous with ICLab's observations. If the page has changed significantly between the snapshot and the observation, the text recorded by the Wayback Machine may not be related to the reason why the page was censored. To estimate how much this affects our results, Figure 1 shows the empirical cumulative distribution function (ECDF) of the median difference between an ICLab observation and a Wayback Machine observation of the same URL. (The ECDF was computed over the *absolute value* of the median differences, but the plot shows signed medians, hence the unusual shape for an ECDF plot. Error bars are +/− one median absolute deviation.) For 75% of the pages in our study, there is a snapshot within three months of each observation, but for the other 25% the gap can be a year or more. Of course, even if the gap is large, the page may not have changed much in that time. In Section 3.3, we investigate the effects of data gaps on LDA-based topic modeling.

Web pages often contain "boilerplate" text, such as copyright notices and navigation menus, that is irrelevant to the page's topic [1, 26]. We mechanically stripped boilerplate from all documents using Trafilatura [2]. LDA requires all documents in the corpus to be in the same language, so, following the approach of Weinberg et al. [26], we translated all the documents word-by-word into English.

2.2 Dynamic Topic Modelling

We analyze the evolution of the set of censored web pages over time using a *dynamic topic model* [4]. Specifically, we use Dynamic Latent Dirichlet Allocation (D-LDA) as implemented in the `gensim` topic-modelling library [16]. D-LDA has two hyper-parameters: the number of topics to divide the corpus into, and the time granularity with which to model time evolution. We selected these via manual experimentation. Figure 2 shows the number of censored URLs



(a) Daily censored URLs. **(b) Weekly censored URLs.** **(c) Monthly censored URLs.**

Figure 2: Distribution of censored URLs in India over time, at daily, weekly, and monthly granularity.

observed by ICLab as a function of three choices of time granularity: days, weeks, and months. With daily and weekly granularity, the number of censored documents changes wildly from one timestep to the next. We found this made the model unstable, so monthly granularity will be used for the remainder of the paper.

LDA tries to make all of its clusters the same size [27], so if there are fewer topics than the hyper-parameter requires, topics that describe many documents will get split up into multiple clusters. A human analyst will see groups of clusters that appear to have the same (or roughly the same) set of salient words (see Section 3.1). On the other hand, if LDA is not given *enough* clusters to work with, then topics that only describe a small number of documents" will get "lumped together," and an analyst will see clusters with salient words that are too generic to identify a topic. To set this hyper-parameter, we started with a large number (100 clusters) and manually reduced it until we found a good tradeoff, at 25 clusters, between split topics and generic lumps. For the remainder of the paper, we will use models set to produce 25 clusters. After manually recombining split clusters, 14 topics are identifiable.

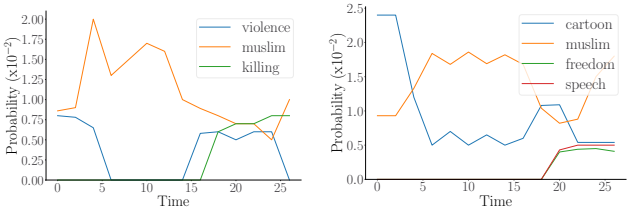
3 EXPERIMENTAL RESULTS

This section presents the results of our evaluation of D-LDA for identifying censored topics (Section 3.1) and for analysing the evolution of censored topics (Section 3.2). We also assess how different variations of the corpus may impact the identification of topics and the evolution of their associated keywords (Section 3.3).

3.1 Identification of Censored Topics

Topic modelling algorithms, in general, associate clusters of documents with a topic, and emit, for each cluster, a set of *salient words* which are characteristic of that cluster's documents. A human analyst must interpret each set of words and assign a more meaningful "label." This step also provides an opportunity to recombine split clusters. Table 1 in Appendix A shows our manually-assigned label and some of the salient words for each of the 14 identifiable topics (plus one "unknown" topic, which we could not interpret).

Sometimes, interpreting salient words for groups of censored documents is straightforward: for example, two of the topics in Table 1 are self-evidently about piracy of music, movies, and TV shows. Censorship infrastructure is often used to enforce copyright law, whether or not that was its original motivation [26]. Several others are clearly about inter-faith conflict and religiously motivated violence, both in India and elsewhere. Religious conflict has been a theme in Indian politics ever since the Partition of 1947.



(a) Riots in India **(b) Danish cartoonist**

Figure 3: Evolution of censored topics over time.

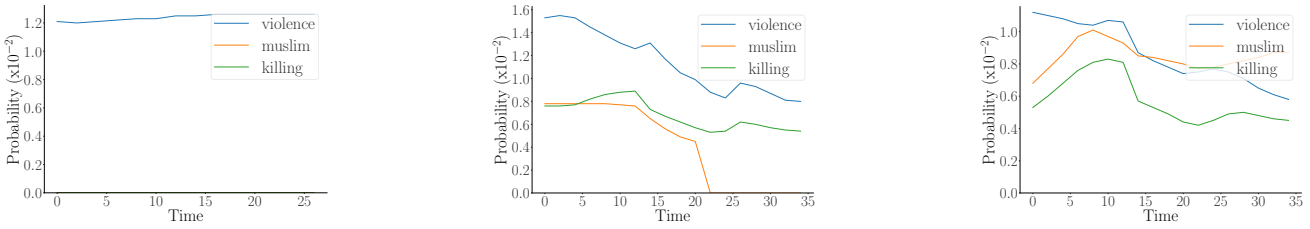
The more difficult topics relate to specific events, only some of which are obviously political. Kurt Westergaard's caricature of the Prophet Muhammad sparked outrage among Muslims in 2005 and discussion of it still appears to be suppressed in India, fifteen years later. Another topic refers to a single (relatively prominent) individual who spoke out against Indian government censorship in 2013, provoking a small online movement. On the other hand, why would "student," "degree," and "education" be salient in documents targeted for censorship? The clue to this topic is "iipm," a.k.a. the Indian Institute of Planning and Management, an unaccredited university which shut down in 2015 after weathering over a decade of accusations of granting worthless degrees and of false advertising.

3.2 Evolution of Censored Topics

An LDA model gives each salient word a likelihood (in the statistical sense) of being associated with *each* topic. The words most strongly associated with each topic are the words listed for that topic in Table 1. D-LDA extends this by computing the likelihood separately at each time step. Therefore, we can indirectly infer changes in censorship policy by observing fluctuations in the level of association between a particular word and a censored topic. Figure 3 shows this fluctuation for three words associated with the "Riots in India" topic and four words associated with the "Danish cartoonist" topic. (The absolute scale of these associations is very small simply because the total probability is spread over thousands of words.)

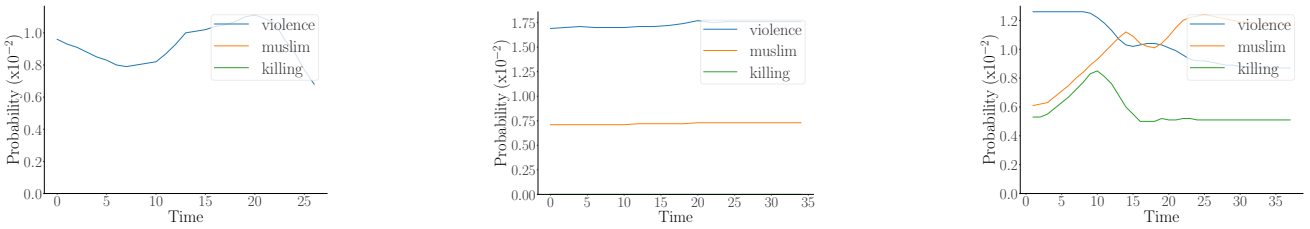
For "Riots in India," the word "muslim" is always salient, but "violence" disappears between months 6 and 14 of the measurement and "killing" does not appear at all until month 16. While we did not have the time to conduct a thorough verification, we suspect this is because there were no actual *riots* between months 6 and 14, and therefore, no reports of violence to be censored. Similarly, perhaps no one involved in such riots was killed until month 16.

The plot for "Danish cartoonist," on the other hand, suggests a change over time in how the controversy was *discussed*. The cartoon itself seems to have been the focus of discussion early on, with broader Islamic issues taking on a greater significance



(a) 25% longest documents. (b) 50% longest documents. (c) 75% longest documents.

Figure 4: Evolution of the probability of keywords for the *Riots in India* topic when filtering the corpus by document size.



(a) Month removed randomly. (b) Month removed every other month. (c) Month removed every two months.

Figure 5: Evolution of the probability of keywords for the *Riots in India* topic when removing data from the corpus.

after five months. "Freedom" and "speech" were not part of the discourse—at least within India—until a year and a half later.

This is all very tentative and ought to be validated by human inspection of the censored documents and cross-referencing with news archives. However, it shows how mechanical topic analysis can bring subtleties of an evolving policy to an analyst's attention.

3.3 Topic Evolution under Corpus Variations

D-LDA is typically applied to data sets with no gaps. As we mentioned in Section 2.1, our data set has gaps because the Internet Archive's observations of censored pages were not concurrent with ICLab's observations of the censorship. ICLab's observations themselves also have gaps, as will be apparent from Figure 2, and we were not even able to analyze ICLab's full data set due to resource limitations. To analyze the effect of data gaps on the identification of topics and the evolution of their defining words, we conducted two experiments in which we introduced *more* gaps into the data set to observe how gracefully the model degraded.

Removing shorter documents from the corpus. We built three alternative D-LDA models based on modified corpora with the shortest 25%, 50%, and 75% of all the documents discarded. As before, we configured D-LDA to output 25 topics. D-LDA produced fewer intelligible topics as the input corpus got smaller. Most of the topics produced by the smaller models are the same as topics found by the full model. However, to our surprise, a few topics (such as "pornography") *only* appear in the smaller models. (The list of identified topics for each model can be found in Appendix B – Table 2.)

Perhaps more interesting is that removal of documents changes the evolution of the association of words with topics. For instance, Figure 4 shows very different trends in the association of "violence," "muslim," and "killing" with the "Riots in India" topic than those seen in Figure 3a. This suggests that small pages (e.g. blog posts) may be important to understand the direction of censorship policy.

Removing monthly data from the corpus. To shed light on how gaps in monthly data (e.g. due to failures in the censorship measurement infrastructure) can affect the results of D-LDA models, we built three more restricted models in which (a) one randomly selected month, (b) every other month, and (c) one out of every three months were discarded from the data set. These models also identified fewer, different topics (listed in Appendix B – Table 2) and estimated topic evolution differently (as shown in Figure 5).

In summary, gaps in data availability are a severe problem for D-LDA. They may cause it to fail to identify important reasons why documents are censored, and they may also cause it to report trends in word association inaccurately. Thus, for providing increased utility, longitudinal censorship measurement platforms should strive to ensure a robust capacity to acquire data, perhaps by deploying redundant measurement nodes inside countries of interest, towards maximizing data availability and enabling accurate topic modeling.

4 CONCLUSIONS

In this work, we applied dynamic topic models to longitudinal censorship measurement data to understand how the textual content of web pages associated with censored topics evolves over time. Our findings show that these models can provide useful insights on the evolution of censorship policies within a given country, but are prone to produce inaccurate results due to gaps in data availability.

Future work. We intend to extend our models to multiple countries, making use of more of ICLab's data, and of data collected by other censorship monitors. We are particularly interested in looking for geographic correlations in the topics censored by different countries. This could shed light on the phenomenon of "censorship leakage" [8] and could also uncover cases where one country's policy has influenced another.

Short of manual labeling (infeasible even at this scale), we are not aware of any principled method for validating whether the topics generated by D-LDA accurately reflect the content of the document corpus. To increase our confidence that the topics generated by D-LDA are not an artifact of either the algorithm or the data collection, we intend to run alternative state-of-the-art dynamic topic models, e.g. DETM [11], and compare the topic clusters they generate with the clusters generated by D-LDA.

REFERENCES

[1] Ziv Bar-Yossef and Sridhar Rajagopalan. 2002. Template detection via data mining and its applications. In *Web Conference*. 580–591. DOI: 10.1145/511446.511522

[2] Adrien Barbaresi. 2021. Trafilatura: A Web Scraping Library and Command-Line Tool for Text Discovery and Extraction. In *Annual Meeting of the Association for Computational Linguistics*. 122–131. DOI: 10.18653/v1/2021.acl-demo.15

[3] Mehrab Bin Morshed, Michaelanne Dye, Syed Ishtiaque Ahmed, and Neha Kumar. 2017. When the Internet Goes Down in Bangladesh. In *Computer Supported Cooperative Work*. 1591–1604. DOI: 10.1145/2998181.2998237

[4] David M. Blei and John D. Lafferty. 2006. Dynamic Topic Models. In *International Conference on Machine Learning*. 113–120.

[5] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research* 3 (March 2003), 993–1022.

[6] Abdelberi Chaabane, Terence Chen, Mathieu Cunche, Emiliano De Cristofaro, Arik Friedman, and Mohamed Kaafar. 2014. Censorship in the Wild: Analyzing Internet Filtering in Syria. In *Internet Measurement Conference*. 285–298.

[7] Le Chen, Chi Zhang, and Christo Wilson. 2013. Tweeting under Pressure: Analyzing Trending Topics and Evolving Word Choice on Sina Weibo. In *Online Social Networks*. 89–100.

[8] Shinyoung Cho, Rishab Nithyanand, Abbas Razaghpanah, and Phillipa Gill. 2017. A Churn for the Better: Localizing Censorship Using Network-Level Path Churn and Network Tomography. In *Conference on Emerging Networking EXperiments and Technologies (CoNEXT)*. 81–87. DOI: 10.1145/3143361.3143386

[9] Alexander Darer, Oliver Farnan, and Joss Wright. 2018. Automated discovery of internet censorship by web crawling. In *Web Science*. 195–204.

[10] Ronald Deibert, John Palfrey, Rafal Rohozinski, and Jonathan Zittrain (Eds.). 2010. *Access Controlled: The Shaping of Power, Rights, and Rule in Cyberspace*.

[11] Adji B. Dieng, Francisco J. R. Ruiz, and David M. Blei. 2019. The Dynamic Embedded Topic Model. (2019). arXiv:1907.05545 [cs.CL]

[12] Arturo Filasto and Jacob Appelbaum. 2012. OONI: Open Observatory of Network Interference. In *Free and Open Communications on the Internet*. 8 pages.

[13] FortiNet. 2005–. FortiGuard Labs Web Filter. https://fortiguard.com/webfilter

[14] Devashish Gosain, Anshika Agarwal, Sahil Shekhawat, Hrishikesh B. Acharya, and Sambuddho Chakravarty. 2018. Mending Wall: On the Implementation of Censorship in India. In *Security and Privacy in Communication Networks*. 418–437. arXiv:1806.06518 [cs.CR]

[15] Derek Greene and James P. Cross. 2017. Exploring the Political Agenda of the European Parliament Using a Dynamic Topic Modeling Approach. *Political Analysis* 25, 1 (2017), 77–94.

[16] Matthew Hoffman, Francis Bach, and David M. Blei. 2010. Online Learning for Latent Dirichlet Allocation. *Advances in Neural Information Processing Systems* 23 (2010), 856–864.

[17] Internet Archive. 1996–. Wayback Machine. https://web.archive.org

[18] Jeffrey Knockel, Masashi Crete-Nishihata, Jason Q. Ng, Adam Senft, and Jedidiah R. Crandall. 2015. Every Rose Has Its Thorn: Censorship and Surveillance on Social Video Platforms in China. In *Free and Open Communications on the Internet*. 10 pages.

[19] Jeffrey Knockel, Lotus Ruan, and Masashi Crete-Nishihata. 2018. An analysis of automatic image filtering on WeChat Moments. In *Free and Open Communications on the Internet*. 12 pages.

[20] Marco Linton, Ernie Gin Swee Teo, Elisabeth Bommes, CY Chen, and Wolfgang Karl Härdle. 2017. Dynamic topic modelling for cryptocurrency community forums. In *Applied quantitative finance*. 355–372.

[21] Arian Akhavan Niaki, Shinyoung Cho, Zachary Weinberg, Nguyen Phong Hoang, Abbas Razaghpanah, Nicolas Christin, and Phillipa Gill. 2020. ICLab: A Global, Longitudinal Internet Censorship Measurement Platform. In *Symposium on Security and Privacy*. 214–230.

[22] Raymond Rambert, Zachary Weinberg, Diogo Barradas, and Nicolas Christin. 2021. Chinese Wall or Swiss Cheese? Keyword filtering in the Great Firewall of China. In *Web Conference*. 472–483. DOI: 10.1145/3442381.3450076

[23] Reethika Ramesh, Ram Sundara Raman, Matthew Bernhard, Victor Ongkowijaya, Leonid Evdokimov, Anne Edmundson, Steven Sprecher, Muhammad Ikram, and Roya Ensafi. 2020. Decentralized Control: A Case Study of Russia. In *Network and Distributed Systems Security Symposium*. 18 pages.

[24] Ram Sundara Raman, Adrian Stoll, Jakub Dalek, Reethkia Ramesh, Will Scott, and Roya Ensafi. 2020. Measuring the Deployment of Network Censorship Filters at Global Scale. In *Network and Distributed System Security Symposium*. 16 pages.

[25] Rima S. Tanash, Zhouhan Chen, Tanmay Thakur, Dan S. Wallach, and Devika Subramanian. 2015. Known Unknowns: An Analysis of Twitter Censorship in Turkey. In *Workshop on Privacy in the Electronic Society*. 11–20. DOI: 10.1145/2808138.2808147

[26] Zachary Weinberg, Mahmood Sharif, Janos Szurdi, and Nicolas Christin. 2017. Topics of Controversy: An Empirical Analysis of Web Censorship Lists. *Privacy Enhancing Technologies* 2017, 1 (2017), 42–61.

[27] Pengtao Xie, Yuntian Deng, and Eric Xing. 2015. Diversifying Restricted Boltzmann Machine for Document Modeling. In *Knowledge Discovery and Data Mining*. 1315–1324. DOI: 10.1145/2783258.2783264

[28] Xueyang Xu, Z. Morley Mao, and J. Alex Halderman. 2011. Internet censorship in China: Where does the filtering occur?. In *Passive and Active Network Measurement*. 133–142.

[29] Tarun Kumar Yadav, Akshat Sinha, Devashish Gosain, Piyush Kumar Sharma, and Sambuddho Chakravarty. 2018. Where The Light Gets In: Analyzing Web Censorship Mechanisms in India. In *Internet Measurement Conference*. 252–264. arXiv:1808.01708 [cs.CY]

APPENDIX

A Labeled Topics

Table 1 sheds additional light on the different topics discovered by the D-LDA model. For each list of salient words discovered by the dynamic topic model, we provide a label for the corresponding topic. We also include an example "Unknown" topic, whose salient words we were unable to make sense of in order to generate a label.

Table 1: Labeled topics and their associated keywords.

Topic	Words
Backlash against URL Blocking	*mahesh, murthy, year, content, writer*
Chinese-language spam	邮箱, 上海货运公司, 上海冷藏运输公司, 传真, *voice*
Danish cartoonist	*cartoon, danish, newspaper, muslim, freedom, religious, speech*
Educational institute fraud	*student, degree, iipm, institution, education, university, campus*
Homicide	*people, kill, brother, human, peace, innocent, sister*
Islam	*muslim, religion, story, domain, book, life, right*
Muslim violence	*state, muslim, violence, illegal, displace, many, woman*
Online streaming services	*video, upload, quality, live, streaming, enjoy, share, hd*
Pirated music/movies	*download, latest, search, music, copyright, movie, inception*
Religion-motivated killing	*people, muslim, victim, kill, religion, police, indian*
Religious websites	*com, website, http, let, wordpress, islampeace*
Riots in India	*riot, killing, indian, people, government, anti, violence*
Rohingya Muslim crisis	*muslim, kill, people, burmese, human, buddhist, refugee*
Saudi Yemen conflict	*saudi, attack, yemeni, force, kill, martyr, member*
Unknown	*people, life, know, read, medium, several, think, war*

B Corpus Variations

Here, we take a closer look on the coherent topics we were able to label (from the 25 clusters of salient words generated by D-LDA) after applying different data removal operations to our corpus.

Removal of documents. Table 2 shows the coherent topics we were able to label when D-LDA was run using different fractions of our corpus. Specifically, we ordered documents by length, and ran D-LDA with the longest 100% (4 577, i.e., all documents), 75% (3 432), 50% (2 288), and 25% (1 144) of the documents in the corpus.

Removal of monthly data. Table 3 shows the coherent topics we were able to label when D-LDA was run with different artificial data gaps. Besides a setting where no documents were removed (yielding 4 577 documents as before), we removed data at different monthly intervals. Specifically, we removed data from one random month (yielding 4 270 documents), from every other month (yielding 1 919 documents), and from two every three months (yielding 1 469 documents).

Table 2: Intelligible topics discovered when D-LDA is run on only the longest N% of the censored documents.

All *4 577 documents*	**Longest 75%** *3 432 documents*	**Longest 50%** *2 288 documents*	**Longest 25%** *1 144 documents*
	A film		
Backlash against URL blocking	Backlash against URL blocking	Backlash against URL blocking	
Chinese-language spam			
Danish cartoonist	Danish cartoonist	Danish cartoonist	Danish cartoonist
Educational institute fraud	Educational institute fraud	Educational institute fraud	Educational institute fraud
Homicide			
Islam	Islam	Islam	Islam
	Mob violence		
Muslim Violence			
Online streaming services		Online streaming services	
Pirated music/movies	Pirated music/movies	Pirated music/movies	
	Pornography	Pornography	Pornography
Religion-motivated killing	Religion-motivated killing		
Religious websites			
Riots in India	Riots in India	Riots in India	Riots in India
Rohingya Muslim crisis	Rohingya Muslim crisis	Rohingya Muslim crisis	Rohingya Muslim crisis
Saudi Yemen conflict	Saudi Yemen conflict	Saudi Yemen conflict	Saudi Yemen conflict
	Terrorism		

Table 3: Intelligible topics discovered when D-LDA is run with some observations erased.

No gaps *4 577 documents*	**One random month erased** *4 270 documents*	**Every other month erased** *1 919 documents*	**1/3 of months erased** *1 479 documents*
Backlash against URL blocking			Backlash against URL blocking
Chinese-language spam			
Danish cartoonist	Danish cartoonist	Danish cartoonist	Danish cartoonist
Educational institute fraud	Educational institute fraud	Educational institute fraud	Educational institute fraud
	Ethnic violence		
Homicide			
Islam	Islam	Islam	Islam
		Muslim	
Muslim Violence			
Online streaming services	Online streaming services	Online streaming services	
Pirated music/movies	Pirated music/movies		
		Pornography	Pornography
Religion-motivated killing			
Religious websites			
Riots in India	Riots in India	Riots in India	
Rohingya Muslim crisis	Rohingya Muslim crisis	Rohingya Muslim crisis	Rohingya Muslim crisis
Saudi Yemen conflict	Saudi Yemen conflict	Saudi Yemen conflict	Saudi Yemen conflict