Journal of Mason Graduate Research
Volume 5 Number 1 Year 2017
© Author
ISSN: 2327-0764
Pages: 17-36

The Geography of Anonymous Communications: Detecting Online Censorship Events

BRIAN SANDBERG

George Mason University

Social media pervades everyday life, and its existence, and the censorship of it, can have a considerable impact on international issues or events. The goal of this research was to detect censorship in anonymity networks and social media platforms. The research aimed to discover complementary evidence of censorship events in anonymous communications and social media activity, including evidence of multinational efforts to collaborate on censorship. To understand censorship patterns and impacts, anomaly detection and similarity methods were developed from Tor client usage metrics and Twitter usage patterns to detect country-level anomalous behavior and to identify similar patterns across multiple countries. Twitter content was analyzed using word embedding techniques to improve ongoing phrase-based data collection methods to facilitate explanation of potential censorship events. This approach successfully demonstrated detection of anomalies in Tor usage including those instances that reflect potential censorship events with similar patterns of censorship across country borders.

Keywords: social media; Twitter; anonymity networks; Tor; anomaly detection; censorship

INTRODUCTION

As content producers, social media platforms dominate the user generated data ecosystem. As content consumers, more than 60 percent of Americans use Twitter and Facebook as news sources (Barthel, Shearer, Gottfried, & Mitchell, 2015). Given the high usage and dependency on social media sites for sharing and consuming information, one wonders what the impact would be if these sites were no longer accessible. Online censorship can take on many forms, but all forms erode free speech and the ability to share information. Governments can censor access to social media sites for various political reasons (Howard, Agarwal, & Hussain, 2011). Social media companies themselves censor content and accounts at their own discretion, marking a fundamental shift in power from government to private corporations by which free speech is limited or protected. For instance, any individual or organization can submit a request to Twitter to have content or accounts removed (Twitter, 2015). While content takedowns by governments or social media companies are concerning (Ammori, 2014; Heins, 2014), the focus of this paper was to detect large-scale censorship, including similar patterns of censorship across multiple countries.

To detect censorship events, I designed anomaly detection and similarity methods leveraging the client usage patterns derived from the Tor network. The Tor Project is the most mature and largest deployed anonymous communication network available (Danezis, Diaz, & Syverson, 2009; Dingledine, Mathewson, & Syverson, 2004). Tor is free software with an open network of volunteer-operated relays that allow users to improve their privacy and security online. It is estimated that Tor has over a million daily users with over half of them in Europe. Tor was originally designed by the Naval Research Laboratory (NRL) to protect military and government communications and today it is used by law enforcement, journalists, activists, dissidents, and many others concerned about privacy. Tor conceals the user's actual location and identity and can be used to circumvent censorship. When oppressive governments block access to social media sites, citizens can leverage Tor to regain access to blocked content. These governments can also blacklist Tor nodes to further suppress access to and sharing of information.

This paper presents detection methods that use Tor client usage metrics and Twitter usage patterns to detect country-level anomalous behavior and identify similar patterns across multiple countries. In addition, Twitter content was collected and analyzed to understand usage patterns, improve ongoing data collection methods, and to help explain potential censorship events. Measuring usage patterns of Twitter per country is generally straightforward. Twitter publishes information about its users and provides a public Application Programming Interface (API) that can be used to estimate location-based usage activity over time. Measuring accurate usage and activity patterns of the Tor network is more complicated. Using aggregate client usage data collected by Tor nodes, reasonable country-level estimates can be obtained. Correlating activities across both platforms can provide unique insights into specific types of events including censorship and political events. Using these methods, I explored two primary questions: (1) Can usage patterns in the Tor network be modeled to detect country-level anomalous behavior and can similar patterns be identified across multiple countries? (2) Can data collection methods for the Twitter platform be enhanced to help detect and explain censorship events?

BACKGROUND AND LITERATURE REVIEW

SOCIAL MEDIA CENSORSHIP

Twitter provides a reporting mechanism that allows individuals, companies, or governments to request content or accounts to be censored or removed. Generally, censor requests are for content that may be illegal or questionable in the respective jurisdiction. These censor requests can include court orders served on Twitter for defamatory statements, Twitter terms of service (TOS) violations, intellectual property or copyright violations, and other legal and non-legal requests. The Twitter transparency reports disclose country-level statistics about censor requests. Incidents that have resulted in censored content are reported to the Lumen Internet censorship database[1].

Lumen is a project of the Berkman Klein Center for Internet & Society at Harvard University. The purpose of the Lumen database is to collect and analyze legal complaints and requests for removal of online material. In a recent Twitter transparency report, Turkey was identified as the country issuing the largest number of censorship requests (Tanash, Chen, Thakur, Wallach, & Subramanian, 2015). Over the period of July 1 to December 31, 2015, there were 450 removal requests from court orders, and 1,761 removal requests from Turkish authorities (Twitter, 2015). The latter involved content based on violations of personal rights and other local laws. Following terror attacks in Suruç, Ankara, and Istanbul, Twitter received requests by the Turkish government to remove content containing images of victims. Related analysis discovered that the number of censored tweets for Turkey is actually two orders of magnitude larger than what Twitter reported (Tanash et al., 2015). This research found that the vast bulk of censored tweets contained political content critical of the Turkish government. This raises the concern that similar trends hold for other countries. Simply relying on censorship reports from social media companies is not adequate.

DETECTING ONLINE CENSORSHIP

Online censorship is prevalent throughout the world. The OpenNet Initiative (ONI)[2] has detected censorship from many countries (Deibert, Palfrey, Rohozinski, & Zittrain, 2008). There are a variety of techniques and mechanisms used for enforcing online censorship or Internet filtering. Most rely on returning a block page to content requests informing users that an attempt

[1] Lumen database https://lumendatabase.org/

[2] The OpenNet Initiative: Collaboration between Citizen Lab at the Munk School of Global Affairs at University of Toronto, the Berkman Center of Internet and Society at Harvard University, and the SecDev Group in Ottawa.

to access a webpage is unsuccessful. Censorship mechanisms also include injecting TCP/IP reset (RST) packets, altering Domain Name Servers (DNS) responses, redirecting traffic through transparent proxies, and using explicit web pages notifying users that content has been blocked (Nguyen & Armitage, 2008).

There also exists a variety of methods to detect or measure these censorship mechanisms. Research has been done on block page detection using web page classification techniques (Dalek et al., 2013) and automatically identifying filtering tools (Jones, Lee, Feamster, & Gill, 2014). Methods to detect the various strategies of enforcing censorship can involve active or passive measurement. Active monitoring uses a target list of destinations, while passive monitoring approaches collect information about users' interaction with services. Passive measurement involves interference tests to measure information about a variety of services, which run on many independent probes deployed at the network edge. Depending on the type of interference to detect, different collection methods are required. For instance, detecting blocking requires reachability information, while detecting performance degradation requires finer grained performance statistics. The research presented in this paper avoids active or passive measures and is driven by a unique approach that uses proxies such as anonymity networks and social media usage patterns to detect censorship around the world. Some research has been conducted on detecting and flagging anomalous events in the Tor network (Danezis, 2011). No work to date was found that addresses similar anomalies across countries or Tor anomalies correlated with social media activity in support of censorship detection. Detection of cross-border censorship is a unique contribution of this work.

CHALLENGES AND RISKS OF CENSORSHIP DETECTION

There are a number of challenges when using anonymity networks and social media platforms as censorship detectors. The censoring of social media sites makes it difficult for citizens to report events via those platforms. Most people are not equipped to circumvent censorship via anonymization tools. Reporting or publicizing censorship events must therefore rely on alternative communication tools or external third parties. Anonymity networks, such as Tor, may also be blocked. Citizens may also fear reporting these events as that may risk their own safety (Bodle, 2013; Chaabane, Manils, & Kaafar, 2010).

There are risks to researchers conducting censorship detection. Researchers in this area should be aware of the risks and ethics involved in detection or measuring censorship, whether running probes that test connections to websites that may be banned or using tools that circumvent censorship. Probing tools do not protect privacy of those running them and measurements are published with IP addresses or other personally identifiable information (PII). Violation of jurisdictional laws and consent for measurement are examples of ethical questions and challenges that have been highlighted when developing methods for censorship measurement (Jones, Ensafi, Feamster, Paxson, & Weaver, 2015).

TOR AND TWITTER AS PROXIES FOR DETECTION

This research aims to detect country-level censorship events of social media sites, as well as the Tor network, which is often used to circumvent censorship. The approach taken in this research models usage behaviors to detect signals of censorship at the country level. Data used from the Tor network included daily client usage metrics at the aggregate country level. Data used from Twitter platform included geo-coded data collections to analyze trends in frequency of country-specific activity and content-based collections to analyze and explain events.

A variety of data challenges exist with Tor and Twitter data that must be addressed to effectively be used to support new techniques to detect censorship events. The geographic dimension of social media streams can be used to support analysis of country-level events. This requires accurate extraction of location information from user generated content. While some users self report their physical location, most users produce and consume content without sharing their physical location. Twitter's geo-based collection methods provide fairly accurate location information, but geo-coded tweets only represent a tiny fraction of the overall traffic (Chandra, Khan, & Muhaya, 2011). Further, the actual content is generally off-topic and noisy, making geo-based data collection ineffective for explaining events.

While geo-based collection methods can be useful for analyzing frequency of tweets and modeling country-level usage trends, content-based collection methods based on keywords, phrases, and hashtags, are useful for event analysis. Unfortunately, content-based collections only contain very small or no relevant posts containing location information. Location can be inferred using metadata such as language, time zone information, profile descriptions, etc (Mahmud, Nichols, & Drews, 2012). The approach used here combined the self reported location information with the inferred location to improve data collections for detecting and analyzing events. For specific countries being monitored for censorship, country-level bounding boxes were used to model and detect changes in normal activity patterns from geo-coded posts over time. Accurately detecting statistically significant changes in country-level tweets can help determine the likelihood of a censorship event.

Content-based methods were used to collect Twitter data to identify and characterize activity over time and to identify communities of interest that contribute to explaining the occurrence of an event. Typically, content filters are defined by the user and applied to the data collection process. To collect highly relevant Twitter content to analyze events, I designed a new approach to automatically identify effective keywords, phrases, and hashtags to use as content filters. A neural network based on word2vec models was used to automatically identify similar word associations relating to content of interest (Mikolov, Chen, Corrado, & Dean, 2013). Word2vec models are used to learn vector representations of words or word embeddings. It is a computationally-efficient predictive model for learning word embeddings from raw text, such as Twitter posts. This automated approach was developed because it is impossible to know a-priori what terms are associated with a censorship event at any particular time.

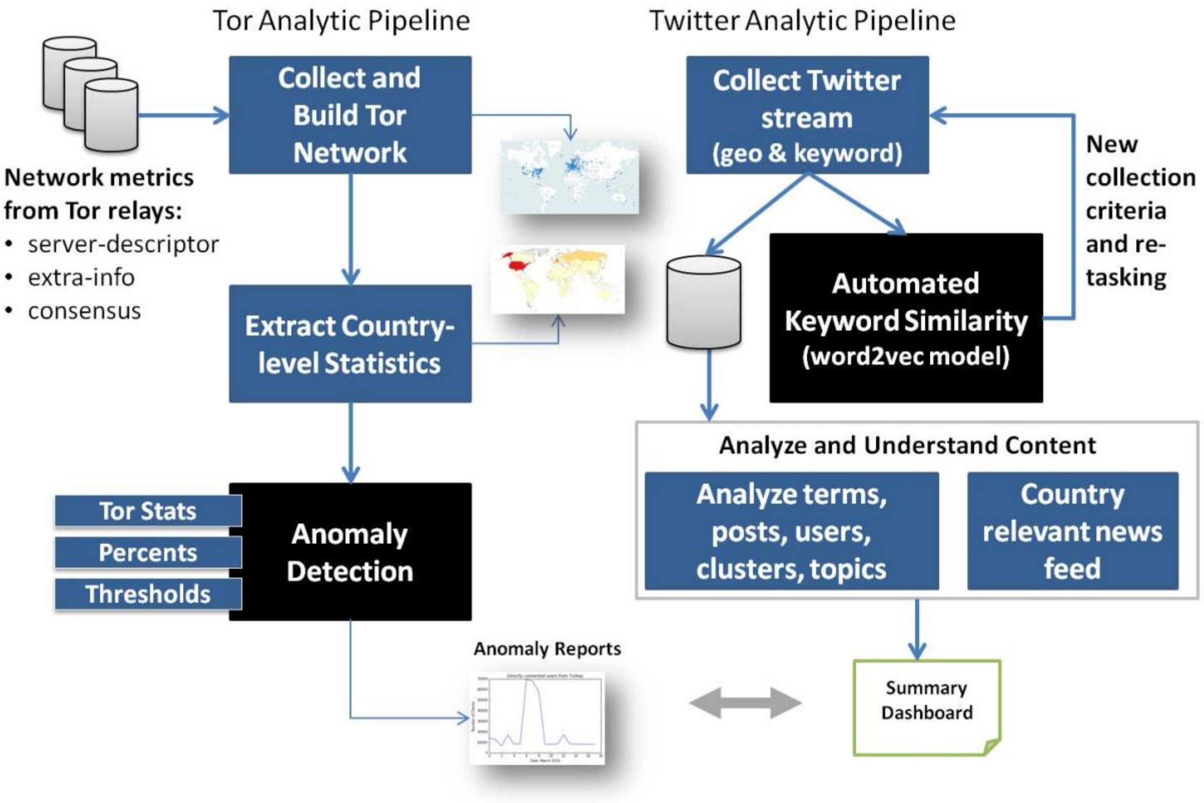
Social media users who also use the Tor network are actively obfuscating their physical location, so place mentions in posts, profile descriptions, and other metadata are the only useful options to extract location information from these types of users (Mahmud et al., 2012). In addition, anomalous activity in Tor client usage patterns can indicate that an event is occurring. These country-level anomalies in Tor usage can be used to task country-specific collections from social media platforms. There are many reasons why anomalies occur in Tor usage, such as file sharing applications and malicious bots (McCoy, Bauer, Grunwald, Kohno, & Sicker, 2008). These types of anomalies are associated with spikes in Tor usage, so are less of interest when detection censorship.

Tor usage patterns vary widely across countries due to population, Internet penetration, familiarity with technology, risk of using technology, among other factors. Because different countries can vary widely in usage patterns, user-specified thresholds were used to more effectively detect events. In this work, I used thresholds to take into account population and usage dynamics and help minimize false alarms. I then used Spatial (point-density) Clustering to detect time series outliers from country-level Tor anomalies. This clustering approach was designed to detect similar patterns of activity across countries.

METHODOLOGY AND RESULTS

Figure 1 outlines the analytic workflow developed in this research. The primary components covered in this paper are highlighted in black boxes. These represent an anomaly detection component that is capable of identifying similar patterns of anomalous activity across different countries in the Tor network and development of a neural network to enhance social media data collection for explaining censorship events. Complete analytic pipelines were developed for Tor and Twitter to support censorship detection.

Figure 1. Analytic Workflow


Tor Analytic Pipeline
Twitter Analytic Pipeline
Collect and Build Tor Network
Network metrics from Tor relays:
• server-descriptor
• extra-info
• consensus
Extract Country-level Statistics
Tor Stats
Percents
Thresholds
Anomaly Detection
Anomaly Reports
Collect Twitter stream (geo & keyword)
New collection criteria and re-tasking
Automated Keyword Similarity (word2vec model)
Analyze and Understand Content
Analyze terms, posts, users, clusters, topics
Country relevant news feed
Summary Dashboard

Anomaly Detection in Tor Usage

Counting individual users, or more specifically individual clients, in the Tor network would reduce the anonymity and safety of the users. Therefore, all information and metrics collected from the Tor Network are at the aggregate level.

Tor Usage Metrics and Anomaly Detection. Tor relay nodes perform data collection services throughout the Tor network. Each relay in the public Tor network performs data collection services and this data is aggregated and made available to the research community[3]. Relays and directory authorities publish relay descriptors so that clients can select relays for their circuits through the Tor network. To count the number of users connecting to the Tor network per country on a daily basis, I acquired Tor metric data from the Tor servers including the Relay Server Descriptors, Relay Extra-info Descriptors, and Network Status Consensus data. These files were collected remotely from the Tor network nodes and stored locally. The Relay Server Descriptors contain information that relays publish about themselves. The descriptor data archive contains one descriptor per file and Tor clients require this information to function properly. The Relay Extra-info Descriptors are self-published like Server Descriptors, but are not downloaded by clients. They contain the client country codes and counts for usage statistics[4]. The Network Status Consensuses is a single document compiled and voted on by directory authorities once per hour, ensuring that all clients have the same information about the relays that make up the Tor network. Directory authorities are special relays that track the overall network. They maintain a list of currently running relays and periodically publish a consensus together with other directory authorities.

The analysis was based on data collected for the entire month of March 2016. Tor metric data that was collected from Tor network nodes was uncompressed and parsed to extract the relevant content to perform usage analysis and anomaly detection. A separate file was generated for each day of the month and each contained information for all nodes in the Tor network (over 7,000 nodes at the time of this research). Table 1 describes each column in the Tor relay files. The DirClients field contained the country codes and counts of client usage for each country. This information was extracted from all files to produce an output file containing country code and usage counts for each day of the month. The sum, mean, and median usage values were computed for each country. In addition, country size (km^2), population, and Internet penetration for each country was added to the output file. This information was necessary for comparing Tor usage with overall Internet usage, population, and Internet penetration. Finally, Freedom House Index indicators for each county was added to compare anomalous events to the degree to which a country was considered free, partly free, or not free. The Freedom House Index is a yearly survey and report that measures the degree of civil liberties and political rights in every nation. The

[3] https://metrics.torproject.org/collector.html

[4] https://collector.torproject.org/recent/relay-descriptors/extra-infos

choropleth maps below show global Tor client usage (Figure 2), global Internet penetration (Figure 3), and Freedom House indicators showing countries not free (red), partly free (dark yellow), and free (light yellow) (Figure 4).

Table 1. Tor Relay Files (extracted fields)

Field	Description
Name	Name of Tor relay node
Fingerprint	Fingerprint for this router’s identity key
Flags	Types of relay node – Middle (M), Exit (E), Guard (G), HSDir (H)
IP	The IP address of the relay node
OrPort	Port at which this Onion Router (OR) accepts TLS connections for the main OR protocol
ObservedBW	Observed bandwidth of relay node (Mbits/s)
GuardClients	Country codes and counts for entry guards based on directory requests
DirClients	Country Codes and counts for director clients based on directory requests (used in the analysis)
Uptime	Relay node up time
Longitude	Longitude of relay node
Latitude	Latitude of relay node

Unexpected increases or decreases in client connections were used to detect anomalies in Tor usage per country. Only a percentage of Tor relays report client connections, and at the time of this research, approximately 33% of the relays provided these metrics. Detecting Tor censorship is based on the number of users per day per country, and the values used were the actual reported numbers by each relay. Since Tor metrics are reported on user activities based on the previous day, detections are based on the previous day’s data. The number of Tor client connections can be similar to the previous day, or they can rise or fall at some level. Large drops can signal censorship of the Tor network, while large increases can signal critical use cases for the Tor network (e.g. censorship circumvention, protest events, malicious activity). As indicated, some of the drops in usage may be attributed to a government blocking the Tor network, as it enables citizens to circumvent censorship. Governments may just block citizen access, but allow continued access by government officials. Detecting these events also requires taking into account the number of users in the target country. Some countries have a very small population of users, so having a small number or zero connections on a given day may not be significant. It is useful to look at usage statistics that take into account population and Internet penetration. In addition, drops in usage can also be a result of Internet outages that are unrelated to Tor.

Figure 2. Tor Client Usage

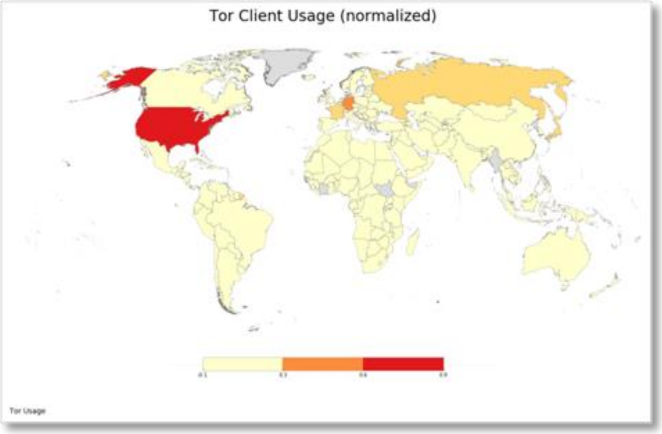


Figure 3. Internet Penetration

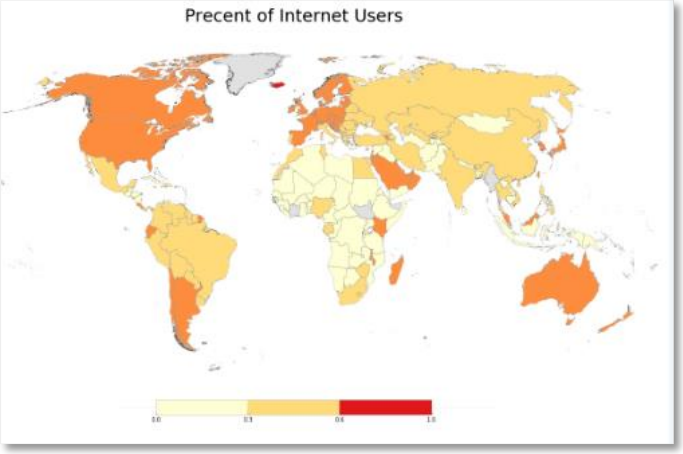
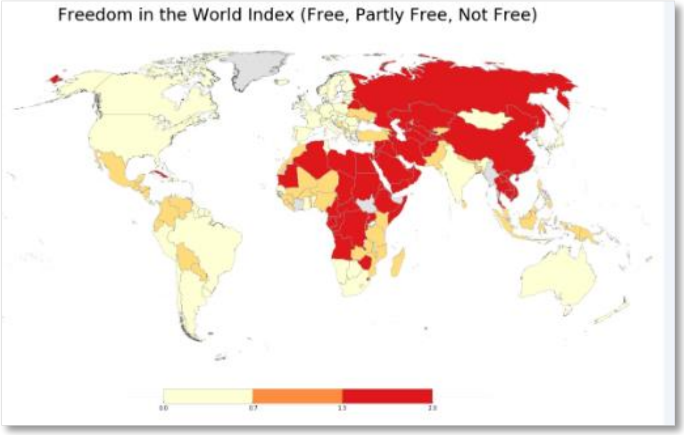


Figure 4. Freedom House Indicators



Anomaly Detection Results. All results were based on data gathered for the entire month of March 2016. I designed the analytics to run daily or monthly and the framework was set up to run experiments on different daily and monthly periods. Table 2 summarizes the top 10 countries by Tor usage over the month of March. The median and mean values are shown for the number of Tor users for each country. If the median values are summed for all countries and divided by four, there was an estimated 1.4M Tor users worldwide per day for the month of March 2016. When Tor relays collect usage data, each country's daily counts are rounded up to the nearest multiple of eight to help protect individual user safety. Correcting by dividing by four provides an average adjustment of the total usage counts. The Detections column in Table 2 shows no anomalies were detected for the top 10 countries.

Table 2. Top 10 Countries by Tor Usage

Country	Median	Mean	Detections
United States	962624	1021603	0
Russia	626432	692093	0
Germany	558968	551529	0
France	309600	312392	0
United Kingdom	235400	235180	0
Italy	154592	152102	0
Spain	147088	143589	0
Brazil	142488	139887	0
Japan	137216	138454	0
Canada	126096	127140	0

Table 3 also summarizes the top 10 countries by Tor usage, but takes into account Internet penetration (mean number of users divided by number of Internet users). This percentage result is shown in the Percent column. In half of these countries, anomalous behaviors were detected in Tor usage. All of these countries were extremely small, except Moldova. Moldova is an interesting outlier based on country size and population. According to ONI, Internet users in Moldova enjoy largely unfettered access to the Internet despite the government's restrictive and increasingly authoritarian tendencies. Past research provided evidence of mounting second- and third-generation controls (Deibert, Palfrey, Rohozinski, & Zittrain, 2008). Moldova is listed at "partly free" by Freedom House.

Table 3. Top 10 Countries by Tor Usage as a Percentage of Internet Access

Country	Size	Population	Internet	Percent	Mean	Detections
Vatican City	<1	842	480	11.88%	57	4
American Samoa	197	54343	3040	10.97%	333	2
Wallis & Futuna	274	15561	1337	8.61%	115	1
Nauru	21	9488	560	4.56%	26	7
Monaco	2	30508	27671	3.71%	1026	0
Republic of Moldova	33843	3583288	1748645	3.18%	55631	2
Anguilla	96	16086	10424	2.66%	278	0
San Marino	61	32742	17000	2.45%	416	0
Gibraltar	7	29185	20660	2.13%	440	0
Turks & Caicos Islands	497	49070	14760	1.95%	287	0

Table 4 summarizes the top 10 countries having the highest number of anomalies detected for the same period. Detections included both increased and decreased usage, which were summed to show total detections. These results were based on user-defined thresholds of 2.0X for usage increase and 0.5X for usage decrease. Experimentation determined these were reasonable thresholds to start with to minimize false alarms. Most detections involved countries with very low Tor usage, so it is likely necessary to vary thresholds based on population or median Tor usage. An interesting outlier based on population and median Tor usage was Oman. According to ONI, Oman engages in extensive filtering of pornographic Web sites, gay and lesbian content, and anonymizer sites used to circumvent censorship (Deibert, Palfrey, Rohozinski, & Zittrain, 2008). Censorship mechanisms in Oman also blacklisted sites that were critical of Islam and

Web sites on illegal drugs (published Aug 6, 2009). Oman was listed as "not free" by Freedom House.

Table 4. Top 10 Countries for Detected Anomalies

Country	Population	Median	Detections
Oman	3286936	3144	18
Tuvalu	10782	0	18
Niue	1190	0	14
Falkland Islands (Malvinas)	2932	32	12
Montserrat	5215	16	12
Norfolk Island	2210	0	12
Christmas Island	1502	8	11
Tokelau	1337	0	10
Tonga	106440	24	9
Central African Republic	5391539	48	8

Similarity Detection and Results. The Density-based Spatial Clustering of Applications with Noise (DBSCAN) algorithm was used to detect similar anomalous patterns over all countries and time period of the study. The data used for clustering were the actual detections previously computed using the daily Tor usage counts for each country ('0' = no detection, '1' = decrease detection, '2' = increase detection). Each row in the data represented a country and each column represented the daily computed detection score for that day and country. There were 31 columns of data representing each day of March and 240 rows representing each country code[5]. The analysis framework can easily be scaled up to multiple months or even multiple years. The goal of clustering was to automatically learn similar usage patterns or anomalous behaviors across multiple countries. A time series of data was labeled for each geographic location. Each location can then be grouped and labeled according to their pattern similarity. These behaviors may relate to specific types of events such as collaborative censorship or a multinational political event.

Spatial clustering decomposes detections into grouping of similar detection patterns. To detect similarity among at least two countries, the DBSCAN instance was created using parameter settings of 0.01 for epsilon and 2.0 for the minimum number of samples. DBSCAN found four

[5] ISO 3166 Country Codes https://dev.maxmind.com/geoip/legacy/codes/iso3166/

clusters using the March detection data. The cluster labeled ‘0’ (yellow markers in Figure 5) is a large cluster that reflects a normal usage pattern of Tor across countries. In other words, these 188 countries did not have any detections of anomalous behavior. The cluster labeled ‘-1’ (white markers) is a smaller cluster of 34 countries with numerous detections, but no similar patterns. The number of detections per country ranged from one to 18 detections. The cluster labeled ‘2’ (green markers) represents two countries with similar detection patterns, Cook Islands and Sierra Leone. While there does not seem to be any interesting outcome to this similarity matching, the cluster labeled ‘1’ (red markers) may be of interest. Cluster 1 involves two countries, China and Bhutan with similar detection patterns. Figure 5 shows the DBSCAN clusters on a map with the red makers highlighting the China and Bhutan cluster.

Figure 5. Clusters of Detected Anomalies.

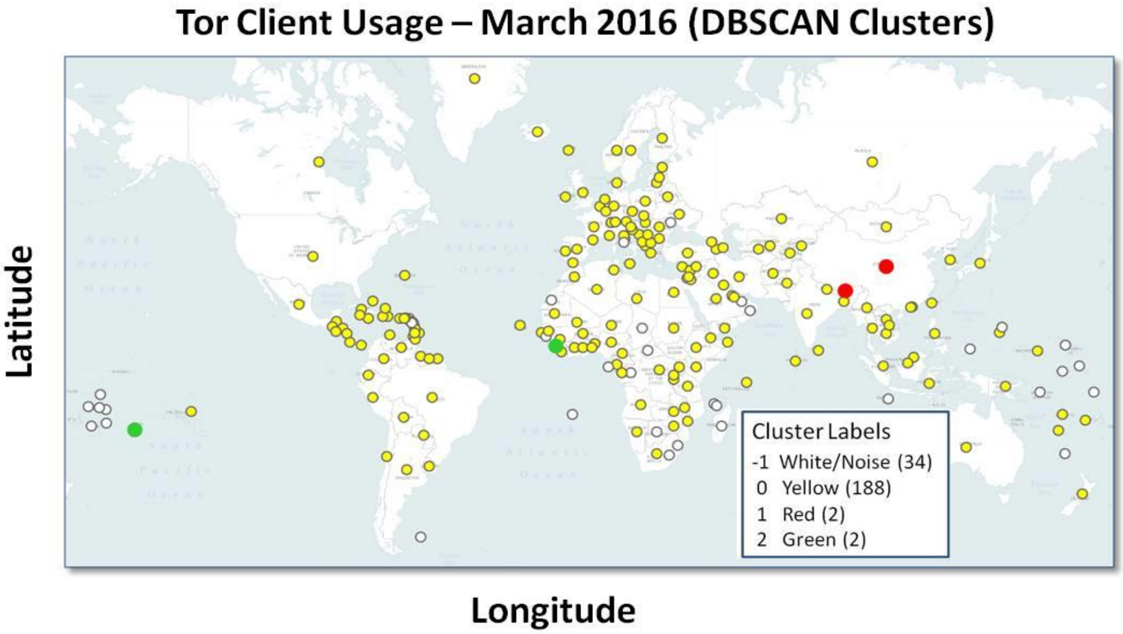
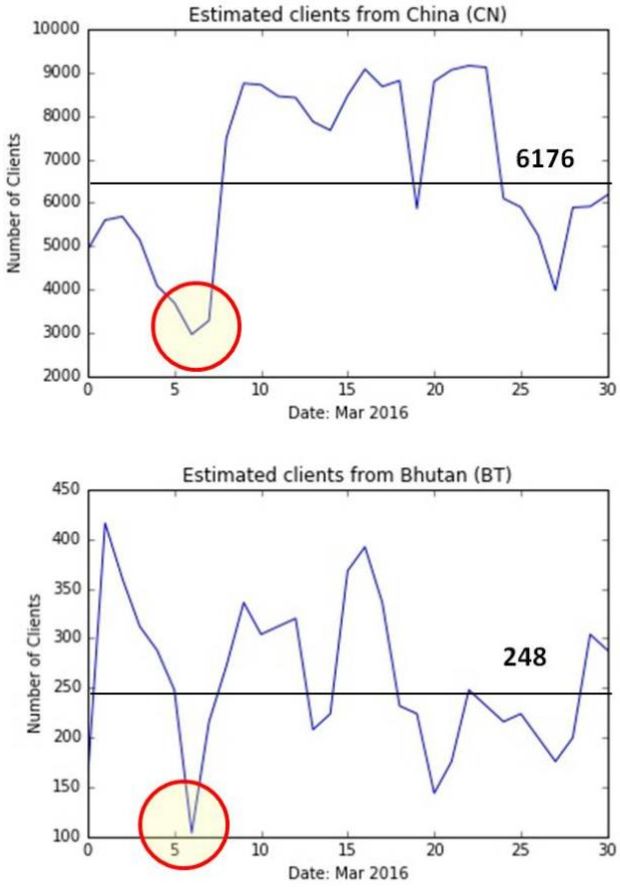


Figure 6 shows the plots of daily Tor activity for both China and Bhutan with red circles highlighting the similarly detected anomaly on March 7th. In both cases, Tor usage was significantly decreased as compared to each country’s median usage indicated along the black horizontal lines.

While this similarity pattern between these two countries may be coincidental, there are a number of interesting aspects to this similarity matching. Both countries share a contiguous border of 470 kilometers with Bhutan to the south of China. Territorial disputes have been a source of potential conflict and they have conducted regular talks on border and security issues. Freedom House ranks Bhutan as “partly free” and China as “not free.” While ONI does not maintain a global internet filtering profile on Bhutan, Freedom House reports that the Bhutan government occasionally blocks access to web sites containing pornography or information deemed offensive to the state. According to ONI, China maintains one of the most pervasive and sophisticated regimes of Internet filtering and information control in the world (Deibert, Palfrey, Rohozinski, & Zittrain, 2008). China blocks access to social media sites including Twitter, Facebook, and YouTube. Among many other blockages, China blocks web sites that discuss Tibetan independence. Bhutan has a strong cultural, historical, religious, and economic connection to Tibet (Bhutan-China Relations, 2016).

Figure 6. Similar Detections in China and Bhutan



Explaining Events with Twitter

Social media sites are increasingly used for breaking news, eyewitness accounts, and organizing events. Popular content can propagate through the network very rapidly. Given that a group of users deem certain information important, that information can flow quickly and can be used as input to event detection. Characteristics of Twitter messages can be analyzed for a variety of insights including sentiment, topic, and event detection. The unique language used by Twitter users and the restricted length of messages limit the use of traditional text mining tools to understand underlying behaviors and events.

Geo vs. Content Based Data Collection. An additional objective of this research was to extract social media events that could help explain anomalies in the Tor network. Detected anomalies may be motivated by political uprisings, censorship, or other types of events. The primary goal was to generate early warning indicators of censorship that could then be used in response planning or reporting. Social media data can compliment Tor usage metrics to support early warning alerts as well as explanation of events. Both require an accurate and automated technique to collect relevant data from Twitter.

Collecting Twitter data using bounding boxes for target countries is straightforward and useful to detect when geo-streams degrade or disappear. The processing requirements to scale this approach to every country with continuous data processing can be expensive. Though, it is effective if the goal is to accurately analyze a few countries to model and detect when rates of geo-streams deviates from a normal baseline. In addition to geo-based collections for activity modeling, content-based collections were used to explain potential censorship events. Setting up Twitter data collections requires careful consideration for capturing user interactions that contain content analysts care about. The selection of the right keywords, phrases, and hashtags to use can significantly improve the relevance of the collection, but human selection of these terms is prone to error and can result in data collections that are not useful to answer analytical questions or model target behavior. The next section describes the approach taken to overcome these limitations.

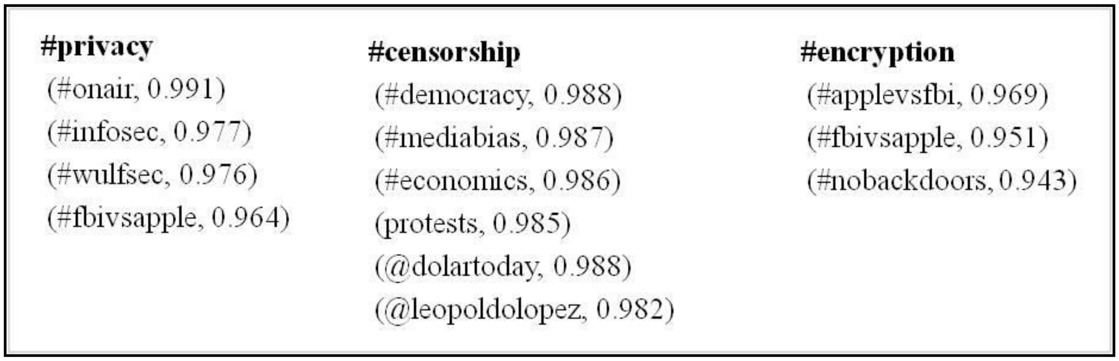
Enhanced Automated Data Collection. While it is difficult and expensive to collect all content, new methods are required to optimize the collection process. To gather real-time and on-topic content about target audience and events, a new natural language processing (NLP) model was developed. This model aimed to automatically produce better terms for filtering the Twitter stream and result in better representations of content for supporting the analysis task. The approach used a neural network trained to reconstruct linguistic contexts of words to improve term selection. As new terms were learned, the collection criteria were updated and the collection service re-tasked.

Specifically, a word2vec model was used to automatically identify similar word associations relating to content of interest (Mikolov et al., 2013). Since training is a high cost compute process, pre-existing models were used. After integrating trained models, they were used to map each word to a vector of several hundred elements, which represent that word's relationship to other words. This approach learned continuous word embeddings from raw text by associating words with points in space. The spatial distance between words describes the similarity between those words. This was represented by a list of words, where each word is also represented by a vector of two dimensions. A displacement vector (i.e. a vector between two vectors) describes the relation between the two words. Comparing the displacement vectors can find pairs of words that have a similar relation to each other. The model resulted in an estimate of the probability of two words occurring near each other.

The word2vec model was run and tested on Twitter data collected over a 48-hour period of March 2-4, 2016 and used three seed terms (i.e. #privacy, #censorship, #encryption). This collection resulted in 19,998 tweets with seed frequency counts as follows: #privacy (2548), #censorship (1016), and #encryption (954). The model estimated which words occurred in adjacent positions in the input text. The result of the word2vec model provided recommendations for new terms for future collections. The new terms for each seed are shown in Figure 7 with the model output probability of the two words occurring near each other. Bold terms are seed terms and list of terms represent closest associations with probabilities.

From these results, example candidate terms for future collections included '#infosec', '#mediabias', and 'protests.' There were also indicators of Twitter users to collect and analyze their networks (e.g. 'wulfsec') and countries to set up a specific bounding box collections (e.g. Venezuela based on newly discovered users '@dolartoday', '@leopoldolopez'). As this test collection was performed in early March of 2016, the seed term #encryption also identified the FBI's efforts to access Apple's operating system ('applevsfbi', 'fbivsapple', 'nobackdoors'). It would be difficult for humans to come up with these new terms, particularly relevant user accounts or specific countries that would help target more useful data collections for explaining events of interest.

Figure 7. Result of word2vec model.



#privacy
(#onair, 0.991)
(#infosec, 0.977)
(#wulfsec, 0.976)
(#fbivsapple, 0.964)

#censorship
(#democracy, 0.988)
(#mediabias, 0.987)
(#economics, 0.986)
(protests, 0.985)
(@dolartoday, 0.988)
(@leopoldolopez, 0.982)

#encryption
(#applevsfbi, 0.969)
(#fbivsapple, 0.951)
(#nobackdoors, 0.943)

DISCUSSION

Results from the censorship detection methods based on Tor usage metrics were very promising, though more work is required to combine social media activity and Tor usage as a more comprehensive censorship detection and analytic framework. Future research will look for complimentary evidence of censorship events by means of connections to Tor and Twitter activity at the country level.

The automated method for enhanced Twitter data collection based on word2vec also demonstrated useful results. Unfortunately, the selected time period for the Twitter data collection did not overlap with the China-Bhutan detection discovered on March 7th. This will require a continuous and iterative approach with Tor detectors and Twitter collections running in parallel.

From an anomaly detection perspective, there is a need to improve detectors by taking population and Tor usage into account. A usage-sensitive detector will avoid noisy detections at lower levels of Tor usage where more random usage behavior occurs. Since Tor is often blocked in a number of countries, such as China, Iran, and Syria, Tor Bridges can be used to provide a way for clients to use Tor even when it is blocked. Data based on Tor Bridges needs to be incorporated into this research going forward. Finally, more analysis is needed to explore the relationship between Tor usage anomalies and the degree to which a country is labeled “free,” “partly free,” or “not free” by Freedom House.

Conclusion

This paper described a methodology for detecting censorship events using Tor Network usage data and a new automated data collection method for Twitter to help explain detected events. Instead of relying on active or passive measurements that require extensive support from in-country participants and potential risks to the safety of the researchers, this approach used Tor and Twitter as proxies for censorship detection and explanation. This research contributed unique analysis into clustering of countries with similar anomalous behavior in Tor usage providing new insights into potential collaborative multinational censorship activities.

References

Ammori, M. (2014, June 20). The “new” New York times: Free speech lawyering in the age of google and twitter. *Harvard Law Review, 127* (2259). Retrieved from https://harvardlawreview.org/2014/06/the-new-new-york-times-free-speech-lawyering-in-the-age-of-google-and-twitter/.

Barthel, M., Shearer, E., Gottfried, J., & Mitchell, A. (2015, July 14). The evolving role of news on twitter and facebook. *Pew Research Center: Journalism & Media*. Retrieved October 19, 2017, from http://www.journalism.org/2015/07/14/the-evolving-role-of-news-on-twitter-and-facebook/.

Bodle, R. (2013). The Ethics of Online Anonymity or Zuckerberg vs. Moot. *ACM SIGCAS Computers and Society*, *43*(1), 22–35. https://doi.org/10.1145/2505414.2505417

Bhutan-China relations (2016) *Wikipedia*, https://en.wikipedia.org/wiki/Bhutan%E2%80%93China_relations

Chaabane, A., Manils, P., & Kaafar, M. A. (2010). Digging into Anonymous Traffic: A Deep Analysis of the Tor Anonymizing Network. In *2010 Fourth International Conference on Network and System Security* (pp. 167–174). https://doi.org/10.1109/NSS.2010.47

Chandra, S., Khan, L., & Muhaya, F. B. (2011). Estimating twitter user location using social interactions–a content based approach. In *2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing* (pp. 838–843).

Dalek, J., Haselton, B., Noman, H., Senft, A., Crete-Nishihata, M., Gill, P., & Deibert, R. J. (2013). A method for identifying and confirming the use of URL filtering products for censorship. *Presented at the Proceedings of the 2013 Conference on Internet Measurement Conference,* ACM, (pp. 23–30).

Danezis, G. (2011). *An anomaly-based censorship-detection system for tor. Tor tech report 2011-09-001*. Retrieved from https://research.torproject.org/techreports/detector-2011-09-09.pdf.

Danezis, G., Diaz, C., & Syverson, P. (2009). Systems for Anonymous Communication. *CRC Handbook of Financial Cryptography and Security*, 1-61.

Deibert, R. J., Palfrey, J. G., Rohozinski, R., & Zittrain, J. (Eds.). (2008). *Access denied: the practice and policy of global internet filtering (information revolution and global politics)*. Cambridge, MA: MIT Press.

Dingledine, R., Mathewson, N., & Syverson, P. (2004). *Tor: The second-generation onion router*. Washington, DC: Naval Research Lab.

Heins, M. (2014, June 20). The brave new world of social media censorship. *Harvard Law Review, 127*, 325. Retrieved from: https://harvardlawreview.org/2014/06/the-brave-new-world-of-social-media-censorship/.

Howard, P. N., Agarwal, S. D., & Hussain, M. M. (2011). When do states disconnect their digital networks? Regime responses to the political uses of social media. *The Communication Review, 14*(3), 216–232.

Jones, B., Ensafi, R., Feamster, N., Paxson, V., & Weaver, N. (2015). Ethical concerns for censorship measurement. *Presented at the Proceedings of the 2015 ACM SIGCOMM Workshop on Ethics in Networked Systems Research, ACM* (pp. 17–19).

Jones, B., Lee, T.-W., Feamster, N., & Gill, P. (2014). Automated detection and fingerprinting of censorship block pages. *Presented at the Proceedings of the 2014 Conference on Internet Measurement Conference, ACM* (pp. 299–304).

Mahmud, J., Nichols, J., & Drews, C. (2012). Where is this tweet from? Inferring home locations of twitter users. In *Sixth International AAAI Conference on Weblogs and Social Media*.

Retrieved from: https://www.aaai.org/ocs/index.php/ICWSM/ICWSM12/paper/view/4605

McCoy, D., Bauer, K., Grunwald, D., Kohno, T., & Sicker, D. (2008). Shining light in dark places: Understanding the tor network. In: N. Borisov, I. Goldberg (Eds.), *Privacy Enhancing Technologies. PETS 2008. Lecture Notes in Computer Science, 5134*. Berlin, Germany: Springer.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. ArXiv:1301.3781 [Cs]. Retrieved from: http://arxiv.org/abs/1301.3781.

Nguyen, T. T. T., & Armitage, G. (2008). A survey of techniques for internet traffic classification using machine learning. *IEEE Communications Surveys Tutorials, 10*(4), 56–76.

Tanash, R. S., Chen, Z., Thakur, T., Wallach, D. S., & Subramanian, D. (2015). Known unknowns: an analysis of twitter censorship in Turkey. *Presented at the Proceedings of the 14th ACM Workshop on Privacy in the Electronic Society, ACM* (pp. 11–20).

Twitter (2015). Removal requests, https://transparency.twitter.com/en/removal-requests.html#removal-requests-jul-dec-2015.